19/prts

INFORMATION-CARRYING AND INFORMATION-PROCESSING POLYMERS

B'

DESCRIPTION

The present invention relates to methods for producing information-carrying polymers and to information-carrying polymers obtained with the following methods: to methods for isolation, amplification and selection of such information-carrying polymers; and to polymeric data storages and DNA-computers, which contain information-carrying polymers, as well as the use of information-carrying polymers for the production of molecular weight standards, as markers or signatures, for the encryption of information, for molecular-scale adhesives, or for the production or processing of miniature molecular structures.

It is well-known to use nucleic acids to mark materials. U.S. Patent No. 5,451,505 describes marking materials with nucleic acids of a length of 20 to 1000bp that may serve the purpose of authentication. However, some basic problems, such as the construction of appropriate sequences or the appropriate coding for the representation of information are not discussed, or even satisfactorily solved.

It is well-known that nucleic acids such as DNA (deoxyribonucleic acid) and RNA (ribonucleic acid) can be used outside of living organisms to process information. Several different approaches for using DNA molecules to process information have been proposed.

WO 97/07440 and [L.M. Adelman, Molecular Computation of Solutions to Combinatorial Problems, *Science*, **266**, 1021-1024, (1994)] describe a method for calculating a graph algorithm ("Hamiltonian path problem") as a decision problem using DNA molecules. The algorithm is implemented by representing the edges and nodes of the graph by DNA sequences. The calculation of the algorithm is performed by generating a set of paths of the graph by hybridizing complementary DNA sequences.

From this set of paths, all paths that are a false solution or no solution of the algorithm are removed by biomolecular methods. Once all steps have been executed successfully, either at least one DNA strain containing the correct solution of the problem is left in the end, or no DNA strain is left, which means that there is no solution for the algorithm.

The above-described method presupposes that enough molecules are available that every possible solution of the respective problem instance occurs statistically at least once in the original set of paths. However, this cannot be guaranteed due to the statistical nature of the hybridization operation. Furthermore, it is possible that cyclical graphs are created in the first step, which increases the number of false solutions.

The approach described in WO 97/07440 shows that symbol processing in vitro with DNA molecules is in principle possible. However, the described approach can hardly be used in practice: One problem is that the algorithm is not deterministic but only stochastic, and the obtained result is only at a certain probability correct. A further problem is that the implementation of the algorithm is not efficient (run-time optimized), so that the calculation is slow.

It is explained that the method used is particularly suitable for the solution of NP-complete problems (a class of problems for which only deterministic solution algorithms with exponential run-time are known, and for which it is assumed that efficient deterministic solution algorithms do not exist), because the algorithm necessitates only a linear run-time due to the parallelism of the executed operations. However, this argumentation is erroneous, because the linear run-time is compensated by an exponential number of molecules. Thus, the exponential nature of the problem remains unchanged.

All in all, the method described in WO 97/07440 is too limited to be useful for molecular information processing beyond the described algorithm: The Hamiltonian path problem is permanently coded, other algorithms cannot be calculated, and every problem instance has to be coded anew. Programming is not possible, and all steps of the algorithm are executed manually. There is no input/output system. All in all, the method is not suitable for programming algorithms or for implementing a computer.

WO 97/29117 and [Frank Guarnieri, Makiko Fliss, Carter Bancroft, Making DNA Add, *Science*, **273**, 220-223 (1996)] describe a method for performing additions with DNA molecules. The addition is carried out as a shift operation with overflow carry in vitro with DNA molecules and primer elongation.

The described method describes additions only. However, even the use as an adder is unfavorable for at least two reasons: Firstly, the addition with the described method is not implemented efficiently (run-time optimized). Moreover, the used system is formally incomplete, because the results obtained by addition cannot be used as numbers for further calculations. Concepts that go beyond the addition are lacking. All in all, the method is not suitable for programming algorithms or for implementing a computer.

[Qi Ouyang, Peter D. Kaplan, Shumao Liu, Albert Libchaber, DNA Solution of the Maximal Clique Problem, *Science*, **278**, 446-449, (1997)] describes a method for implementing an algorithm for solving the "max-clique problem." The algorithm is derived from graph theory and belongs to the NP-complete problems. Like [L.M. Adelman, Molecular Computation of Solutions to Combinatorial Problems, *Science*, **266**, 1021-1024, (1994)], the method can calculate only permanently coded solution algorithms. Also in this case, the result obtained is only a certain probability. The method described by the authors contains no continuing concepts (for example regarding the implementation of other algorithms). All in all, the method is not suitable for programming algorithms or for implementing a computer.

[Eric Winfree, Xiaoping Yang, Nadrian C. Seeman, Universal Computation via Self-assembly of DNA: Some Theory and Experiments, *Proceedings of the 2nd DIMACS Meeting on DNA Based Computers, Princeton University, June 20-12*, (1996)] discusses the implementation of regular grammars in DNA by linking oligonucleotides with "sticky ends." However, the discussion is purely theoretical and fails because of substantial, hitherto unsolved problems. In particular, it is not shown how the sequences necessary to implement grammars have to be constructed. Yet this is the crucial problem that has to be solved in order to implement grammars. The reason for this is that sequences that represent the variables and terminals of a grammar have to be unambiguous as well as sufficiently dissimilar to one another. Moreover, they have to fulfill certain structural, chemical and physical conditions. Otherwise, failed hybridizations among sequences are inevitable, which lead to undesired chain extensions and chain break-off during polymerization, thus making proper functioning of the method impossible. This problem is aggravated exponentially with the number of necessary sequences.

The described method is rejected by the authors themselves, or referred to as insufficient insofar as it does not allow very interesting calculations to be done (see Eric Winfree, Xiaoping Yang, Nadrian C. Seeman, Universal Computation via Self-assembly of DNA: Some Theory and Experiments, *Proceedings of the 2nd DIMACS Meeting on DNA Based Computers, Princeton University, June 20-12*, (1996), page 8, second paragraph, line 1).

Further experiments by the authors are therefore not based on regular grammars and linear polymers, but on context-sensitive grammars for the generation of DNA lattice structures [Eric, Winfree, Furong Liu, Lisa A. Wenzler & Nadrian C. Seeman, Design and self-assembly of two-dimensional DNA crystals, *Nature*, **394**, 539-544, (1998)]. The approach described by the authors is possibly useful for the generation of DNA based nano-structures (fabrication of catalysts, etc.). However, for information processing, the approach is not suited, as the generated lattice structures tend to be troublesome.

The authors raise the possibility of realizing the mathematical concept of "Wang tiles" with the generated lattice structures and possibly using it for calculations, but they merely mention this possibility, without describing how it could be used in the context of information processing.

It is doubtful whether the approach described by the authors is suitable at all for information processing: The lattice structures prevent the amplification of the information-carrying sequences (for example by cloning or PCR = polymerase chain reaction), and make it impossible to use the created structures as data, since they cannot be read out anymore, for example by PCR. Consequently, the selected approach in its current form presents conceptually no possibility to use the created structures in the context of information processing, for example as data.

From the conventional technology, no method is known that makes it possible to use DNA molecules for the implementation of efficient algorithms. No method is known to execute calculations automatically (for example in form of a molecular computer). Furthermore, no method is known for implementing programs in the form of regular grammars with molecular methods, such that

a) different (ideally any desired) regular grammars can be implemented;

b) words of a language that have been generated by regular grammars can be read out;

c) the words of a language that have been generated by regular grammars can be further used for technical purposes (such as information processing, polymer chemistry).

The problem solved by the present invention is to present a method with which it is possible to carry out information processing on molecular basis with regular grammars, without being limited to one or a few grammars. The method should be compatible to conventional computers and allow information processing that is partly molecular-based and partly based on conventional computers. The method should be programmable and it should be possible to largely automate the method. The polymers generated in the method should be readable and usable for further applications and methods.

This problem is solved by a method for manufacturing information-carrying polymers, including:

I. defining a regular grammar G = (∑, V, R, S) with a finite terminal alphabet ∑, a finite set of variables V, a finite set of rules R, and a start symbol S;

II. the NFR method (Niehaus-Feldkamp-Rauhe method) for producing monomer sequences (oligomers or polymers);

III. implementing, with the NFR method, a grammar defined in Step I, by producing with the NFR method monomer sequences that unambiguously represent the set of rules R of a grammar G;

IV. assembling, from the monomer sequences produced in Step III, for each rule of the set of rules R of G an oligomer (*algomer*) representing that rule (*algomer assembly*);

V. linking the oligomers (*algomers*) assembled in Step IV to information-carrying polymers (symbol polymerization).

In the framework of the present invention, the following definitions and abbreviations are used:

algomer — double-stranded oligomer that represents a rule of a given grammar. Algomers can be linked with one another to logomers.

readout PCR — PCR that is used to read out the information contained in the logomers.

Biochip — carrier of a number of nucleic acids that are used to detect complementary sequences; also referred to as microarray, DNA array, gene array, or gene chip.

bit polymerization — process of concatenating algomers together into logomers, when there are only two different elongators.

byte — information unit consisting of 8 bits, or molecule representing 8 bits.

elongator — algomer with two overhang sequences; can ligate with a terminator or an elongator and leads to chain elongation.

grammar — formalism describing languages. The formalism is based on a formal theory of languages [Chomsky, N., Three models for the description of language, *JACM*, **2:3**, 113-124, (1956)], [Chomsky, N., On certain formal properties of grammars, *Inf. and Control*, **2:2**, 137-167, (1959)], [Chomsky, N., Formal properties of grammars, *Handbook of Math. Psych.*, **2**, 323-418, (1963)]

A grammar G describes a language L(G), the alphabet of this language and its syntax. With a grammar G, all words of that language can be generated.

A grammar G is a quadruple ($\sum$, V, R, S) with a terminal alphabet $\sum$, a set of variables V, a start symbol S and a set of rules R.

logomer — polymer that carries symbolic information and that has been generated by the linking of algomers. Correspondingly, a logomer is made of repeating units of algomers. A logomer represents a word of a language L(G), that is generated by a corresponding grammar G.

monomer	single molecule. A plurality of monomers can be linked to long chains, and thus form oligomers and polymers. In the case of deoxyribonucleic acid, the nucleotides (adenine, cytosine, guanine, thymine, as well as analogous bases such as hypoxanthine, etc.) are the monomers
multibyte	an arbitrary data structure that is made of a plurality of bytes.
oligomer	short-chain molecule made of repeating units (monomers). Also, short double-stranded molecules are referred to as oligomers.
PCR	polymerase chain reaction; method for the exponential amplification of DNA. PCR needs a DNA template to be amplified and two primers that start in opposite directions in the template and serve as the starting points of a DNA polymerization. By iterative repeating of melt-hybridization-polymerization cycles, the DNA template is multiplied.
polymer	long-chain molecule made of repeating units (monomers).
rule	also: production rule, substitution rule, or derivation rule. Describes the substitution of symbols by other symbols. Symbol chains can be generated by repeated application of rules.
sequence	Information theory: Series of symbols. Chemistry: Series of covalently bonded monomers. Molecular biology: Series of covalently bonded nucleotides.
complementarity	two sequences are complementary when they can hybridize with one another. For example, in the case of DNA, the sequences `5'att3'` and `5'aaat3'` are complementary, and the sequence `5'acgt3'` is complementary to itself.
start symbol	variable within a grammar, starting from which a symbol chain can be generated by application of the rules of the grammar.
symbol polymerization	the process of concatenating algomers into logomers.
terminal	symbol of a grammar. A terminal cannot be substituted any further.

Terminals are the "letters" of the words of a language L(G).

terminator algomer with one overhang sequence. Can ligate only with an elongator. Leads to chain break-off in a symbol polymerization.

uniqueness unambiguity of sequences among one another. A sequence S of monomers is 10-unique to a set M of other sequences, if no partial sequence of S of a length 10 appears in any other sequence of the set M. The uniqueness can also be given in percent. For example, two sequences of length 20, whose longest common partial sequence is five monomers, are 6-unique to one another, and their uniqueness in percent is (1 – 5/20)*100 = 80%.

variable symbol of a grammar. According to a rule of a grammar, a variable can be substituted by terminals, variables or combinations of terminals and variables.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a view of algomers as described in the explanation of Step IV of the method of the present invention. The algomers each have a double-stranded core sequence, that represents a terminal of a given grammar and at least one single-stranded overhang sequence, which represents a variable of the given grammar, so that an algomer represents exactly one rule of the grammar. X and Y are not variables, but overhang sequences, that can be used for cloning, for example. The letters marked with bars represent complementarity. According to the definition, A0A and A1A are elongators, XsA and AeY are terminators. The algomers shown here represent the grammar for the generation of binary random numbers described below.

Fig. 2 is a view of symbol polymerization as described in the explanation of Step V of the present invention. Algomers are linked into logomers by concatenation (in the case of DNA: Hybridization and ligation). The logomer Xs01010101eY contains a terminated bit series that can be amplified by the overhangs X and Y in a definite orientation.

Fig. 3 is a view of a pattern illustrating the symbol polymerization of binary random numbers after gel electrophoresis and coloring (tracks 1 to 3). Because of the random length of the binary random numbers, the ladder pattern is regular. Track 4 shows a 50bp molecular weight standard. (Gibco BRL, Life Technologies, Catalogue No. 10416-014).

Fig. 4 is a view of cloning of a logomer obtained by symbol polymerization in a vector, as described in the following for the amplification and isolation of logomers.

Fig. 5 is a view of a diagram schematically illustrating a band pattern that is obtained by reading out a binary logomer by PCR and subsequent gel electrophoresis (described below). In the illustrated example, the algomers have a length of 30bp (taking overhang sequences as 1/2). For logomers with known length, it is sufficient to read only the 0's or 1's. Reading both kinds of bits is good for control. The method can also be used for polyvalent logomers (with more than two elongators).

Fig. 6 is a view of a band pattern of a gel electrophoresis after PCR to read out three different logomers. The logomers have been obtained in accordance with the grammar for random numbers of any length, which is described below. The random numbers, read from bottom to top, are: a = 262, b = 97, c = 329. M (track 5) is a 50bp molecular weight standard (Gibco BRL, Life Technologies, Catalogue No. 10416-014).

Fig. 7 is a view of a diagram schematically illustrating the reading out of logomers by restriction digestion as described below. The elongators carry restriction cut sites, that are arranged asymmetrically, so that the restriction digestion of logomers results in an unambiguous cutting pattern of fragments. The cutting pattern corresponds to bands of various lengths, and can be made visible by gel electrophoresis. R1 and R2 are different restriction enzymes, and x and y denote the length of the fragments obtained after the restriction digestion. An x:y ratio of 1:2 is suitable.

Fig. 8 is a view of band patterns that are obtained by gel electrophoresis after the reading out of logomers by restriction digestion. Tracks 7 and 8 contain molecular weight standards (track 7: 50bp ladder, Gibco BRL, Life Technologies, Catalogue No. 10416-014; track 8: 10bp ladder).

Fig. 9 is a view of a schematic representation of a polymeric data storage based on algomers and symbol polymerization as described below. Algomers can polymerize at anchor molecules (AM) bonded to a solid carrier (C). Writing (W) is carried out by a repeating (ReW) of hybridization-ligation-restriction cycles (Hyb, Lig, Res). The obtained logomers can be separated and read out by denaturing or restriction (Den/Dig).

Fig. 10 is a view of a simplified diagram showing the components of a DNA desktop computer as described below. The components are:

A: oligo-synthesizer, B: thermal cycler, C: reaction chambers, D: pipetting device, E: Gel, F: scanner, G: control computer

The numerals denote:

1: addition of oligonucleotides, 2: addition of synthesized oligomers in reaction chambers of the thermal cycler, 3: addition of solutions and molecules needed for algomer assembly, symbol polymerization, for isolating logomers and for reading out.

Fig. 11 is a view of an encryption of logomers as described below. If the terminators and the primers priming therein are unknown, then the corresponding logomer is encrypted and cannot be read out (A). If, on the other hand, a primer priming in a terminator is available or the sequence of a terminator is known, then the corresponding logomer can be read out (B).

Fig. 12 is a view of a Y-shaped molecule that can be used as a terminator for the asymmetric encryption of logomers. Elongators can be linked to the end marked with the 2, and further for example Y-shaped molecules to the ends marked 1 and 3. Linking several Y-shaped molecules with one another, tree-like structures are obtained.

Fig. 13 is a view of logomer with the terminators s and e that are realized as tree-like structures.

Fig. 14 is a view of a logomer V with tree-like terminators that is linked with a vector V. The terminators are configured such that only one branch each of the terminator can be linked with the vector.

Fig. 15 is a view of marking of nucleic acids with logomers as described below: In order to mark genetically engineered or modified products, a logomer is applied using recombinative techniques in a not-transcribed area, e.g. before the promoter (P) of a gene (G) to be marked.

Fig. 16 is a view of marking of documents with logomers as described below.

Track 1 shows the used molecular standard (50bp, Gibco ZBRL, Life Technologies, Catalogue No. 10416-014). Track 2 (0 bits) and track 3 (1 bits) show the band pattern of the readout of logomer No. 330 by PCR from an aqueous solution. Tracks 4 and 5 are the same as tracks 2 and 3, but here the logomer No. 330 is not read from an aqueous solution but was introduced to the readout PCR in form of paper scraps (of about 1mm^2) of 10^9 molecules/µl dried on paper (3M PostIt, dried for 1 hour),

Fig. 17 is a view of production of molecular weight standards by readout PCR, containing a unary logomer as a template, as described below. 1 shows the arrangement of nested primers in the readout PCR, 2 shows the band pattern obtained after gel electrophoresis of the PCR fragments.

Fig. 18 is a view of gluing of surfaces with nucleic acids as described below. C denotes the surface to be glued, Log denotes the logomers that are bonded to the surfaces for gluing. 1 shows the behavior of areas with non-complementary logomers and 2 shows the behavior of areas with complementary logomers.

Fig. 19 is a view of gluing of surfaces with logomers, antibodies and ligands as described below: C_0 and C_1 denote the surfaces to be glued, which can be of the same or of different materials. Logomers (Log) are applied on the surfaces to be glued. Proteins, such as antibodies of the type Ab A and Ab B can bond to the logomers, wherein Ab A and Ab B can be the same or different. Ab A and Ab B are bonded to one another by a ligand (Li).

Fig. 20 is a view of gluing of surfaces with proteins, such as antibodies, without using logomers. C_0 and C_1 denote the surfaces to be glued, which can be of the same or of different materials. Proteins of the type Ab A and Ab B bond directly to the surfaces to be glued, and are bonded to one another by a ligand (Li).

Fig. 21 is a view showing that when linking sequences to longer sequence chains, there may be violations of the required uniqueness, which may lead to failed hybridizations and thus to disfunctionality. This is caused by the generation of new base sequences (marked in the diagram) due to linking.

Fig. 22 is a view of an example of the linking of terminals with a variable. The four illustrated rules of a grammar with a variable A result in four different paths, that overlap over the length of the variable sequence A. Furthermore, some of the sequences for multiple terminals (b, c) overlap, so that three paths converge on the left and on the right.

Fig. 23 is a view of showing that if more than four different variable sequences (A, B, C, D, E) have transitions to the same terminal sequence (0), then at least one base sequence has to be used several times. The dotted frame indicates the iteration step, at which a violation of the uniqueness has to be tolerated, in order to be able to translate all rules of R.

Fig. 24 is a view showing parallel filling in of two sequence elongations for the variables A and B. The dotted frame indicates the iteration step, at which the starting base sequences for the path search are located. From the next iteration step onward, violations of the uniqueness can be tolerated, if necessary. Note that this results in branching beyond group borders (e.g. of the terminal b to the variables A and B).

Fig. 25 is a view of a diagram of a molecule representing 1-byte. The section marked by x contains an unambiguous sequence that represents a predetermined byte value (the 256 nucleic acid sequences needed for the representation of all byte values are listed in the sequence listing). The section marked s contains an unambiguous sequence, that codes the byte position of the corresponding byte within multibytes and serves as a template for PCR reactions (see sequence listing). The sections marked o and e are for the production of multibytes from single bytes and are used as templates for PCR reactions (see sequence listing).

Fig. 26 is a view of a schematic representation of four 1-byte molecules with different byte positions ($s_0 - s_3$). The single bytes can be connected to multibytes.

Fig. 27 is a view showing that for amplification, individual byte molecules can be cloned in genetic vectors (plasmids).

Fig. 28 is a view showing an example of the construction of a byte-representing DNA molecule. The molecule includes several functional sub-units: For X, there are 256 unambiguous base sequences, which represent all values of a single byte. S is an unambiguous sequence that represents the position of a byte (namely, the subsequent X in the strand) within multibytes. O and E are used as unambiguous recognition sequences for the concatenation of single bytes to multibytes.

DNA bytes can be used for labeling can be non-concatenated or concatenated. Cutting out X or SX sub-units yields DNA sections that are used for the "spotting" of biochips. The resulting biochips are used for the reading out of single bytes or of multibytes.

Fig. 29 is a view showing concatenation of bytes to multibytes. In this example, four bytes are linked to a 32-bit data structure. Moreover, the sequences L and R at the ends can function as adaptors, in order to introduce the 32-bit DNA data structure into another DNA. They can carry either specific recombination sites or restriction sites. An example is the labeling of a plasmid in Fig. 31.

Fig. 30 is a view of a diagram showing the configuration of a 32-bit molecule, that has been produced by the linking of four 1-byte molecules. The molecule can be admixed or tacked for labeling to substances to be marked. For the labeling of nucleic acid constructs and genes, the sequences denoted L and R can carry restriction sites or recombination sites, with which they are connected to the molecule to be marked.

Fig. 31 is a view showing labeling of a plasmid with a 32-bit molecule. The 0th byte has the value 109, the first byte has the value 67, the second byte has the value 35, and the third byte has the value 192. As an unsigned 32-bit, the byte patterns of the marked plasmid corresponds to the number 3223536493.

Fig. 32 is a view of a diagram of a 1-byte biochip, that has been spotted with all 256 x-fragments (see Fig. 25 to Fig. 28) from x_0 to x_{255} (X-chip). Only some of the 256 different sequences are shown. To read out multibytes, the corresponding single bytes are preamplified by PCR (s to e) and hybridized individually on separate chips. Thus, four PCR reactions and four chips are necessary for a 4-byte molecule (see Fig. 30).

Fig. 33 is a view of a diagram of a 1-byte biochip, that has been spotted with all 256 x-fragments (see Fig. 25 to Fig. 28) from s_0x_0 to s_0x_{255} (X-chip). Only some of the 256 different sequences are shown. In contrast to the chip type of Fig. 32, the hybridization conditions can be selected such that bytes can be detected in dependence of their position in a multibyte. In this example, only s_0x_i, but no s_nx_i with $n \neq 0$ are detected. Similarly, it is possible to manufacture 1-byte chips that detect only s_1x_i, etc.

This makes it possible to manufacture chips, that can read out multibytes directly (without PCR) and completely. An example of this is the 4-byte chip shown in Fig. 37.

Fig. 34 is a view of a layout of a 1-byte chip. If the chip is implemented as an X-chip (see Fig. 32), then it contains 256 spots with all 256 x-fragments (see sequence listing). If it is implemented as an SX-chip (see Fig. 33), then it contains 256 spots with all s_ix-fragments. Each x-fragment represents exactly one byte value ($x_0 = 0$, $x_1 = 1$, ..., $x_{255} = 255$). The sequences for s and x are selected such that they have melting temperatures that are as similar to one another as possible, but sequences that are as different from one another as possible, in order to preclude failed hybridizations.

Fig. 35 is a view showing manufacturing of multibyte chips from 1-byte SX-chips. This example shows the manufacturing of a 4-byte chip of one SX_0 chip, one SX_1 chip, one SX_2 chip, and one SX_3 chip. Multibyte chips can be arranged to (a) linear or (b) two-dimensional byte arrays.

Fig. 36 is a view of a read out a 32-bit molecule (see Fig. 30 and Fig. 31) with a 4-byte SX-chip (see Fig. 35). For the readout, 32-bit molecules can be hybridized directly with the entire chip. Thus, the 32-bit value can be read out directly (marked in the diagram). Moreover, bytes can be amplified independently from one another and hybridized separately. Alternatively, a multibyte-chip can be made of identical 8-bit units, that are spotted only with X-fragments. In order to maintain the position information of the bytes, the individual bytes must be spotted separated from one another in the corresponding sectors.

Fig. 37 is a view of a multibyte array of identical 1-byte chips (X-chips). Multibyte arrays of X-chips can also be used to read out marking molecules as described above. In addition, multibyte arrays can be used for storage and for the optical display of computer data.

The Sequence Listing shows molecules that are needed for the production of the 32-bit molecules described below. The molecules are assembled to algomers, as described below. The o, s, e, and x units are assembled to bytes, as described below. These are, in turn, assembled to multibytes (in this example to 4-byte

= 32-bit molecules). The 32-bit molecules are then used for labeling and marking and are also used for the manufacturing of biochips (as described below).

Explanation of I: Selection of Grammars

A grammar is a quadruple G = ($\sum$, V, R, S) with a finite terminal alphabet $\sum$, a finite set of variables V, a finite set of rules R and a start symbol S. All words of a language L(G) that is described by G can be formed as words of the terminal alphabet $\sum$, by derivation from the start symbol S according to the rules R.

The definition of a grammar G according to Step I of the method of the present invention is free in that any finite set of terminals and variables can be coded. However, preferable for the present invention is the definition of grammars that allows the binary representation of data, in particular of grammars with

$\sum := \{0, 1, s_0, s_1, s_2, \ldots, s_{n-1}, s_n, e_0, e_1, e_2, \ldots, e_{m-1}, e_m\}$ with $n, m \in \mathbb{N}$,

and $n, m \geqq 0$.

This allows the manufacturing of binary logomers.

For illustration, the following illustrates a regular grammar, defined in accordance with the present invention, for generating random numbers of arbitrary finite lengths in binary notation:

The grammar G = ($\sum$, V, R, S) has a finite terminal alphabet $\sum := \{0, 1, s, e\}$, a set of variables V := {A}, a start symbol S and a set of rules

R :=

{

S := sA

A ➔ 0A

A ➔ 1A

A ➔ e

}

wherein

s := start

e := end.

With this grammar, words of the terminal alphabet

∑ := {0, 1, s, e}

can be formed. All words of the language L begin with “s” and end with “e”, and in between there is a random number (including zero) of random bits (“0” and “1”).

Words of the language L that have been formed according to R are for example:

a) S → sA → se

b) S → sA → s1A → s1e

c) S → sA → s0A → s00A → s001A → s0010A → s0010e

d) S → sA → s1A → s10A → s100A → s1000A → s10000A → s100000A → s100000e

The binary patterns generated as words of the language L can be read as arbitrary but unambiguous representation of data types, e.g. as letters, numbers, alphanumeric characters or character strings. Reading them as the binary representation of numbers, then the above words are in decimal representation:

a) Ø {empty}

b) 1

c) 2

d) 32.

Explanation of II: The NFR method for manufacturing monomer sequences

The NFR method (Niehaus-Feldkamp-Rauhe method) is a method for manufacturing monomer sequences (oligomers and polymers, e.g. nucleic acids) that ensures that the monomer sequences manufactured with this method:

a) have an unambiguous and unique sequence of monomers;

b) are as dissimilar to one another as possible, and therefore if possible do not undergo failed hybridizations;

c) have certain structural properties, such as a certain ratio among the different monomers, the containing or non-containing of certain partial sequences and a maximum matching (homology) among one another.

d) have certain chemical properties, such as the occurrence of certain chemically active partial sequences that influence chemical reactions, in the case of nucleic acids in particular the interaction with proteins (protein binding locations, restriction cut sites, stop codons).

e) have certain physical properties, such as a certain melting temperature, for example.

The NFR method allows the production of unambiguous monomer sequences that are as dissimilar as possible and have predefinable, structural, chemical, and physical properties. It is suitable for the production of monomer sequences for controlled chemical reactions, such as are necessary for molecular information processing, for example. It is used in Step III of the inventive method for the implementation of regular grammars.

The NFR method is an inventive improvement of the Niehaus method for the generation of unambiguous sequences that are as dissimilar as possible, which is described in [Jens Niehaus, DNA Computing: Assessment and Simulation, *Master Thesis Paper at the Faculty of Computer Science of Dortmund University, Department XI*, 116-123, (1998)], which is incorporated by reference herein.

The NFR method is an inventive improvement of the Niehaus method in that only the NFR method allows the manufacturing of monomer sequences as are necessary for the production of monomer sequences in vitro, for example to implement grammars. The reasons for this are:

a) The Niehaus method describes only the construction of sequences from base sequences of a length of 6. Since the length of basic frequencies describes the maximally possible overlaps between different monomer sequences, and thus directly influences the hybridization behavior of sequences as well as the success of the inventive Steps IV and V, the method has to be generalized for base sequences of any length, which is achieved by the NFR method.

b) The Niehaus method allows only the generation of sequences of equal length, whereas the NFR method can generate sequences of different lengths and different numbers. This last aspect is essential for correct hybridization behavior of sequences as well as the success of the inventive Steps IV and V.

c) The Niehaus method ensures unambiguity and maximum dissimilarity of sequences only for sequences that have been generated by a one-time application of the method. However, it is imperative that compatible, that is, unambiguous and maximally dissimilar sequences can be generated also for existing sequences. In contrast to the Niehaus method, the NFR method can generate compatible (that is, unambiguous and maximally dissimilar) new sequences for any given sequence. Herein, the method can be applied any number of times to the same set of sequences.

d) The Niehaus method limits the maximum length of common partial sequences, however it does not limit their number. Therefore, two arbitrary sequences constructed according to the Niehaus method may contain several common partial sequences, which can unintentionally lead to a high homology (sequence matching). The homology then has direct influence on the hybridization behavior of sequences and the success of the inventive Steps IV and V. The NFR method, on the other hand, also carries out a homology comparison during the construction of sequences and thus avoids unintentional failed hybridizations.

e) The Niehaus method does not allow the integration of certain prescribable sequences and partial sequences. However, such sequences are absolutely necessary for the execution and control of certain chemical reactions, such as controlled enzymatic interaction. In the case of nucleic acids, examples of such sequences are protein binding locations, restriction cut sites and stop codons. On the other hand, the NFR method allows the integration of any sequence and partial sequence into the production of monomer sequences and simultaneously ensures unambiguity and maximum dissimilarity.

f) The Niehaus method does not provide the possibility to generate sequences with certain structural, chemical, or physical properties. On the other hand, the NFR method includes the possibility of generating sequences with certain structural, chemical, and physical properties. This also includes the ratio among different monomers and the melting temperature. This is also a necessary condition for the implementation of grammars in vitro.

g) The Niehaus method does not allow a direct generation of rule-representing sequences, as are necessary for the implementation of grammars. On the other hand, the NFR method allows the generation of symbol-representing sequences, which can be linked to rule-representing sequences that are necessary for the implementation of grammars.

h) Only with the NFR method, sequences can be constructed such that the properties of unambiguity and maximum dissimilarity as well as structural, chemical, and physical properties are also valid for partial sequences. For example, in the case of nucleic acids, monomer sequences can be manufactured such that the entire sequence as well as for example the first third of the sequence has a 50% content of GC. This property is a necessary condition for example for the successful concatenation of sequences that are designated to several hybridization events per molecule (such as for example the algomers described below). Only in this manner can the sequences be constructed, for example, such that the different hybridization events per molecule have equal probability.

i) Only the NFR method allows the direct, automatic production of monomer sequences, for example, with an oligonucleotide synthesizer.

j) When linking the sequences together to sequences of longer chains, the uniqueness may be violated. For the correct manufacture of polymers by linking oligomers (such as the linking of algomers to logomers), it is absolutely necessary that the uniqueness violations only occur in a predefined area, so that no uncontrolled failed hybridizations occur (see Fig. 21). This problem occurs in general and must be solved in particular for the linking of algomers and the linking of variables and terminals, because otherwise, a translation of grammars into molecules is impossible. The Niehaus method does not solve this problem and is therefore not suited for the translation of grammars into molecules. The problem is solved by a "parallel extension" strategy, using the NFR method. This is explained in the following.

Monomer sequences are produced by constructing and synthesizing them with the NFR method. To manufacture n monomer sequences, the NFR method is carried out as described in the following, using the following abbreviations:

A := alphabet of the potency $a \in \mathbb{N}$;
in case of DNA, $A := \{a, c, g, t\}$ and $a = 4$.

S_i := sequence.
series of elements of the alphabet A, which corresponds to a sequence of monomers.

l_{seq} := length of sequences to be constructed per method cycle.

S_{seq} := sequence of the length l_{seq}.

$S_{seq,k}$:= $(k\text{-}1)^{th}$ monomer of the sequence S_{seq} ($k \in \mathbb{N}$; $k \leqq l_{seq}$).

l_{bas} := length of the base sequences used for the construction of sequences per method cycle.
$0 < l_{bas} \leqq l_{seq}$.

S_{bas} := sequence of the length l_{bas} = base sequence; sequences of the length l_{seq} are constructed from base sequences.

l_{ov} := maximum length of chain of subsequent monomers that are shared by two sequences generated by the method ("overlap").

l_{ov}/l_{seq} := ratio of maximally allowed sequence repetition to total length of sequence;
1st measure for probability of failed hybridization.

l_{bas}/l_{seq} := ratio of base sequence length to total length of sequence;
2nd measure for probability of failed hybridization.

M_{seq} := set of Sequences.

M_{bas} := set of all base sequences of length l_{bas}.

M_{nobas} := subset of the set of all base sequences of length l_{bas} that is obtained by decomposition from M_{seq}.

$|M_{nobas}|$:= potency of M_{nobas} = number of base sequences in M_{nobas}.

n := number of sequences to be constructed per method cycle.

n_{tot} := total number of sequences produced in all method cycles.

n_{max} := maximum number of sequences that can be constructed per cycle.

h := homology. Measure for the matching between two sequences.

$0 \leqq h \leqq 1.$

t_m := melting temperature. For a DNA sequence, the melting temperature is determined according to the *nearest neighbor* method (see Breslauer, K.J., Frank, R:, Blocker, H., Marky, L.A., Proc. Natl. Acad. Sci., **83**, 3746-3750, 1989):

$tm = \Delta H/(\Delta S + R \text{ x } In(C/4)) - 273.15°C.$

Herein, ΔH and ΔS are enthalpy and entropy of the DNA helix, R is the molar gas constant and C is the concentration of the DNA sequence.

The production method is carried out as a series of any arbitrary but finite number of method cycles, in which n_{tot} sequences are constructed, and a subsequent synthesis, in which all n_{tot} sequences are generated in vitro.

In accordance with the definitions listed above, the maximum number n_{max} of sequences that can be constructed per method cycle can be determined by

$$nmax = f(a, lbas, lseq) := \lfloor(1/2*((a^{lbas})- (a^{lbas/2})-|M_{nobas}|)) / (l_{seq}-l_{ov})\rfloor.$$

This means it is possible to construct maximally n_{max} sequences of the length l_{seq} from the elements of an alphabet A of size a (in the case of DNA, a = 4), that match in maximally l_{ov} contiguous chains of monomers. The number of sequences actually obtained per method cycle can be smaller when certain physical, chemical, or structural properties are desired, or when further sequences are produced in addition to already existing sequences.

In particular, the following method steps are necessary:

1 n_{tot} sequences are constructed in z method cycles and collected in a set M_{seq}.

2 Prior to the method cycle, any unambiguous sequence can be added once to the

sequence set M_{seq}. This can be useful in order to add certain sequences, which should be contained in the set of manufactured monomer sequences because of the further application. It is recommended that not too many and not too long sequences are added, because the method does not guarantee the same properties for these sequences as for the sequences constructed in the following.

3 Start of the method cycle is a view of The structural, chemical, and physical properties that the sequences to be produced must fulfill are defined. The structural properties are at least containing or non-containing of certain partial sequences (if necessary, also the positions, at which the partial sequences in the sequences to be manufactured should be contained or not contained) and the ratio of the number of different monomers to one another (in the case of DNA: ratio GC/AT). The chemical properties are at least the occurrence of certain partial sequences that influence interaction with their own or other substances (in the case of DNA e.g. certain protein binding sites, restriction cut sites such as 5'gatatc3' for EcoRV, the stop codons 5'tca3', 5'tta3', 5'cta3', etc.). The physical properties to be determined include at least the melting temperature t_m of the sequences to be manufactured.

4 The number of $n \leq n_{max}$ of sequences desired per cycle is selected in accordance with the above formula l_{seq} of the sequences to be constructed is selected, and l_{bas} of the base sequences necessary for construction is selected. l_{seq} and l_{bas} are chosen in accordance with the above-mentioned formula such that the number n of sequences to be constructed can be reached. Since l_{bas}/l_{seq} is proportional to the probability of failed hybridization, l_{seq} and l_{bas} are chosen such that the ratio l_{bas}/l_{seq} is as small as possible (values below 0.3 are preferable) and the values for l_{seq} guarantee good hybridization conditions. For DNA, any values for $l_{seq} > 0$ are possible, depending on the temperature.

5 A set M_{bas} of all sequences of the alphabet A of length l_{bas} is generated. The sequences generated in this manner are called *base sequences*. There are always a^{lbas} base sequences. The base sequences are used to construct the sequences to be generated. Each base sequence is assigned one of the states "used" and "unused." Initially, all base sequences are marked as unused.

6 Self-complementary base sequences of M_{bas} are now marked as used. There are exactly $a^{lbas/2}$ base sequences that are self-complementary.

7 If there are already sequences in the set M_{seq}, then sequences compatible thereto can be constructed or the sequences of a subset of M_{seq} can be elongated. For this, in a decomposition process, a set M_{nobas} of all partial sequences with the length l_{bas} prescribed in Step 4 is formed of the sequences already available in M_{seq}:

```
Set $M_{nobas} = \{ \}$.
Every sequences $S_{seq}$ of $M_{seq}$ is now decomposed in $(l_{seq} - l_{ov})$
        decomposition steps.
Start with $i = 0$ with the decomposition step:
While $i < (l_{seq} - l_{ov})$:
        Form a new sequence $S_{new}$ as a sequence of the
        length lbas, including the monomers $S_{seq,i}$ to $S_{seq,i+lbas}$:
        $S_{new} := S_{seq,i}, \ldots, S_{seq,i+lbas}$ .
        Add $S_{new}$ to the set $M_{nobas}$.
        Set $i := i + 1$.
```

The resulting set M_{nobas} is by definition a subset of M_{bas}.

8 All base sequences of the set M_{bas} that also occur in M_{nobas} are marked as used, so that for the construction of sequences exactly $(a^{lbas} - a^{lbas/2} - |M_{nobas}|)$ unused base sequences are left over, of which maximally

$$n_{\max} = \left\lfloor \frac{\left(\frac{\left(a^{lbas}\right) - \left(a^{lbas/2}\right) - |Mnobas|}{2} \right)}{\left(l_{seq} - l_{ov}\right)} \right\rfloor$$

different new sequences can be constructed that have no base sequences or complements thereof in common with one another and with already present sequences.

9 From the base sequences, a directed graph is constructed whose nodes represents certain base sequences: The graph contains exactly a^{lbas} nodes, of which

already ($a^{lbas/2}$ + $|M_{nobas}|$) are marked as nodes. Each node is associated with a base sequence $S_{b(i)}$, that is not complementary to itself (in the following, nodes and their associated base sequence are not distinguished).

An edge from $S_{b(i)}$ to $S_{b(k)}$ exists, if the l_{ov} last letters of $S_{b(i)}$ correspond to the l_{ov} first letters of $S_{b(k)}$, that is, if:

$S_{b(i),2}$,..., $S_{b(i),lbas}$ = $S_{b(k),1}$,..., $S_{b(k),lbas-1}$.

The node $S_{b(k)}$ is called the *successor node* to $S_{b(i)}$.

The node that is associated with the complement of the base sequence that is coded by the node $S_{b(i)}$ is called the *complementary node* to $S_{b(i)}$.

Sequences of length l_{seq} are found by looking for a path with ($l_{seq} - l_{ov}$) nodes. No node may be used more than once. Moreover, for each node that is in one of the paths, the complementary node may not occur in any path. The starting node of the path carries, like the sequence, a state of “marked” or “unmarked.”

Sequences are constructed from the graph by the following method:

```
Mark all unused nodes of the graph as unused starting nodes.
    For every s between 1 and n:
        As long as there are nodes $S_{b(k)}$, that are not yet marked as used starting nodes:
        Select an unused node $S_{b(k)}$.
        Mark node $S_{b(k)}$ as used starting node; furthermore:
        Mark sequence $S_{b(k)}$ and its complement as used.
        Now a new sequence $S_{new}$ is constructed, that is a new element in $M_{seq}$ or elongates an existing sequence $M_{sub}$ of $M_{seq}$:
        Set $S_{new,o} := S_{b(k),1}$.
        If constructed as new sequence, set $i := 0$.
        If constructed as elongation of an existing sequence $S_{sub}$, then set $i := l_{sub}$.
        While $i < l_{seq} - (l_{ov} - 1)$ and $i \geqq 0$:
            If no unused successor node $S_{b(m)}$ to node
```

$S_{b(k)}$ exists, then mark node Sb(k) and its complementary node as unused and set i := i-1.

Else select randomly an unused successor node $S_{b(m)}$ of node $S_{b(k)}$ and mark b_m and its complementary node as unused. In addition, set: i := i + 1, k := m, and $S_{new,i}$:= $S_{b(m),1}$.

If i = l_{seq} – (l_{ov} – 1), then the sequence s_{new} is finished: it includes the letters $S_{new,0}$ to $S_{new,(lseq-lov)}$.

11 The structural, chemical, and physical properties of each of the sequences obtained in Step 10 are determined in comparison to the requirements defined in Step 3. If a sequence does not conform with the preset requirements, then it is deleted, and the nodes and complementary nodes used by it are again marked as unused.

12 If the number of constructed sequences is not sufficient, the method cycle can be repeated, as often as necessary, from Step 3 or 4 onward. In particular, it is possible to set the structural, chemical and physical criteria and the values for n, l_{seq} and l_{bas} new for each method cycle. Thus, it is possible to construct sequences with different structural, chemical, and physical properties, as well as different length, that are still compatible, which means unambiguous and maximally dissimilar.

Else, go to the next step, the in vitro synthesis.

13 The constructed sequences are synthesized in vitro, and in the case of nucleic acids, preferably using an oligonucleotide synthesizer (e.g. ABI 392, ABI 398, ABI 3948 by Perkin-Elmer Applied Biosystems). For this, the sequence data are communicated to the control unit of the oligonucleotide synthesizer and the sequences are manufactured as single-stranded nucleic acids. The sequences can also be ordered commercially. PAGE-purified oligonucleotides (e.g. obtainable at ARK Scientific GmbH Biosystems, 64293) are suitable.

In order to translate grammars into molecules, the above-mentioned problem of uniqueness violation when linking sequences must be solved (see Fig. 21). This problem cannot be solved with the Niehaus method, because the Niehaus method does not include the linking of sequences. In particular, the Niehaus method is not potent enough for the solution of the following partial problems:

- The linking of sequences can lead to a uniqueness violation (see Fig. 21).
- When linking variables and terminals, the sequence of a variable can link to more than one terminal sequence per end and vice versa (see Fig. 22 and Fig. 24), depending on the grammar.
- When linking e.g. several terminals with the same variable, uniqueness violations can occur, that cannot be resolved by a simple path search (see Fig. 23).

The problem is solved by a strategy referred to as “parallel extension”, using the NFR method. More specifically, this is done as follows (illustrated for the construction of variables to predetermined terminals, see Fig. 22 to Fig. 24):

1. A group of sequences (here: the terminals, e.g. sequences for representing bits or sequences with certain chemical or biological properties) is given.
2. Collect for each variable all pairs of terminals whose sequences will frame the variable sequence in the logomer and group them together, one group for each variable (see Fig. 22 and Fig. 24). Arrange the paths of the terminal pairs such that a gap remains between each pair, and the length of the gap corresponds to the variable path lengths ($l_{seq} - l_{bas} + 1$) plus $l_{bas} - 1$ for each of the two transitions. Herein, a transition is the path of the length $l_{bas} - 1$ that connects terminals and variables that belong together. In Fig. 21, the transition consists of the marked base sequences, for example. Transitions can converge (e.g. the terminals a, b c converge to the variable A in Fig. 22) or branch (e.g. the variable A branches to the terminals b, c and d in Fig. 22).
3. Select the last base sequence of each terminal sequence on the left side as the starting node for the corresponding path (see Fig. 24).
4. Generate the terminal-variable-transitions simultaneously, that is, find, in each iteration step, for all paths a successor node whose last monomer is the same for all

paths of a group. If in this step no unused and allowed node can be found, then backtracking is triggered. If this is not possible, then the translation of the grammar into algomers fails at this point and the translation must be repeated with modified (preferably less restrictive) parameters. However, the multiple use of a node in an iteration step is permissible for the transitions that belong together (see Fig. 23 and Fig. 24).

5. Generate also the variable sequences simultaneously as an elongation of the transitions.
6. Generate the variable-terminal-transitions analogously as described in Fig. 4. Herein, the successor cannot be selected freely, but is given by the first $l_{bas} - 1$ nucleotides of the right terminal sequence. The path search follows this condition, in order to find out, whether the paths can be completed.

This method, referred to as "parallel extension" makes it possible to create an interface between a computer and the steps executed in vitro, starting with the synthesizing of oligonucleotides, and to completely automate the steps from the definition of the grammars up to the manufacturing of molecules in vitro.

Explanation of III: Implementation of regular grammars with the NFR method

The production of monomer sequences to represent the set of rules of a grammar in Step III of the method of the present invention is carried out using the NFR method (Niehaus-Feldkamp-Rauhe method) according to Step II of the method of the present invention. To implement a grammar, exactly 2r monomer sequences are produced with the NFR method for r rules of a grammar G, so that the monomer sequences for the s represented symbols (terminals and variables) and the rules R of the grammar G are unambiguous and as dissimilar to one another as possible, and have the demanded structural, chemical, and physical properties.

For this, the monomer sequences are produced according to the NFR method, so that they can be assembled to algomers, as described below. Two monomer sequences each are assembled to an algomer, which represents exactly one rule R of a grammar G. The two monomer sequences both contain sequences that contain the symbols (terminals or variables) that belong together and are required by rule R.

The design of the oligomers according to Step III of the method of the present invention depends on the chemical properties of the employed monomers. In the preferable case of oligomers made of nucleotides,. the following procedure is carried out:

Algomers are double-stranded and have a 5'-strand (*upper strand*) and a 3'-strand (*lower strand*). Upper strand and lower strand can be the link for, respectively, one terminal sequence and one variable sequence.

X, Z are any variables, S is a variable functioning as the start symbol, and y and z are terminal symbols. Then, for each rule of the form:

a) S := yX

an algomer is constructed contains an unambiguous double-stranded core sequence y, to which an unambiguous single-stranded overhang sequence X is appended at the 3' end; ":=" in this case means that S and yX are identical, that is, yX functions as the starter molecule.

b) X ➔ yZ

an algomer is constructed that, that contains an unambiguous double-stranded core sequence y, to which an unambiguous single-stranded overhang sequence X is appended at the 5' end, and an unambiguous single-stranded overhang sequence Z is appended at the 3' end. X can be identical with Z.

c) X ➔ z

an algomer is constructed, that contains an unambiguous double-stranded core sequence z, to which an unambiguous single-stranded overhang sequence X is appended at the 5' end.

Algomers of the form S := yZ and X ➔ z are called *terminators* ("start", "end"), because they lead to chain break-off during polymerization in the linking step V according to the present invention, which is explained below; algomers of the form X ➔ yZ are called *elongators*, because they lead to a chain elongation during the polymerization in the linking step V according to the present invention.

Preferably, the monomer sequences produced in accordance with Steps II and III of the method of the present invention include nucleotides, in particular ribonucleotides, and most preferably deoxyribonucleotides. The monomer sequences constructed in Step II of the method of the present invention preferably contain for example certain sequences, such as stop codons, recognition sequences for enzymes cleaving nucleic acids (restriction nucleases), recognition sequences for proteins binding nucleic acids, such as described in [J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning, A Laboratory Manual, (1989)], [Rolf Knippers, Molekulare Genetik, *Georg Thieme Verlag*, (1997)] and [Benjamin Lewin, Genes V, *Oxford University Press*, (1994)].

Explanation of IV: Assembly of algomers

The assembly of the sequences synthesized in Step III is carried out in Step IV of the method of the present invention. In the context of the present invention, all techniques for assembling oligomer sequences known to a person skilled in the art can be applied to carry out Step IV. In the case of using nucleotide sequences, which is preferable in accordance with the present invention, the following procedure is suitable:

a) The single-stranded sequences belonging to elongators are phosphorylized. For this, several protocols known to a person skilled in the art are possible, for example the following: In a 20 µl preparation, 16 µl of a synthesized 100 µM sequence, 2 µl of a ligation buffer (e.g. 50 nM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 10 mM dithiothreitol, 1 mM ATP, 25 µg/ml BSA Bovine Serum Albumin, New England Biolabs) and 2 µl PNK (polynucleotide kinase, e.g. by New England Biolabs, Catalogue No. #201S or #201L) are incubated for one hour at 37°C. The volumes, incubation time and incubation temperature can be varied in a manner known to a person skilled in the art.

b) In one preparation, the single strands (upper and lower strand) belonging to one algomer are hybridized to one complete algomer. For elongators, it is possible to simply use the phosphorylation preparations of upper and lower strand. For the hybridization, several protocols are possible; it is preferable to use a denaturation step of about 95°C at the beginning (which also deactivates the PNK in the preparations of the elongators) and a slow hybridization. For example, for oligomers of length 30, it is possible to use the following protocol, which can be varied in accordance with the knowledge of a person skilled in the art:

In a 40 μl preparation, 20 μl of the upper strand (100 μM) and 20 μl of the lower strand (100 μM) are heated in a thermal cycler for 5 minutes to 95°C, incubated for 5 minutes to 72°C, and then cooled at 1°C per minute for 25 minutes.

Explanation of V: Manufacture of logomers by concatenating algomers: Symbol polymerization

In Step V of the method of the present invention, algomers are linked to information-carrying polymers (logomers). Since the algomers represent a set of rules R of a grammar G, all generated logomers correspond to words of the language L(G), that is described by the grammar G.

The regulated process of concatenating algomers to logomers is referred to as "symbol polymerization", and is also called "bit polymerization" if the terminal symbols represent bits, that is, if:

$\Sigma := \{0, 1, s_0, s_1, s_2, \ldots, s_{n-1}, s_n, e_0, e_1, e_2, \ldots, e_{m-1}, e_m\}$ with $n, m \in I\!N$,
and $n, m > 0$

In the context of the present invention, all techniques for linking oligomers into polymers that are known to a person skilled in the art can be applied to carry out Step V. In the application of oligonucleotides, which is preferably in accordance with the present invention, it is suitable to incubate the algomers to be linked under the presence of ligase. For example, it is possible to follow the following protocol, which can be modified in accordance with the knowledge by a person skilled in the art:

In a 27 μl preparation, 1 μl of 50μM "start" algomer, 1 μl of 50μM "end" algomer, and 2x each of 10 μl of 40μM elongator algomers (obtained by Step IV of the method of the present invention), 1.5 μl of 10mM rATP and 3.5 μl of T4 DNA ligase with 400 NEB units/μl (e.g. Catalogue No. #202S and #202L NEB) are incubated between 4°C and 25°C for 24 hours.

Other reaction volumes work analogously. Incubation time and incubation temperature can be modified as known to the person skilled in the art.

Fig. 3 shows the result of such a symbol polymerization.

In principle, it is possible to link any number of algomers that are not terminators. For each individual logomer, the polymerization process comes to a standstill when the corresponding molecule carries an algomer at its respective ends, that functions as terminator molecule ("start," "end").

Isolation and amplification of logomers by cloning

A further object of the present invention is a process for isolating and amplifying information-carrying polymers obtained by one of the previous methods by ligating the information-carrying polymers obtained in Step V in cloning vectors, transforming competent cells with these vectors and selecting the successfully transformed bacteria by selection markers.

The logomers obtained from Step V of the method of the present invention as described above (symbol polymerization) can be cloned for isolation and amplification.

For this, the used cloning vector is cut open by restriction enzymes, and a logomer is ligated into it. This method ensures that only completely polymerized logomers (provided with terminators) can be ligated with the target vector. Logomers that are correctly ligated into the target vector again form a ring-shaped molecule. The isolation of logomers is carried out by the transformation of bacteria with the molecule mixture that is the result of the ligation. The vectors carry selection markers (preferably antibiotics resistance, such as ampicillin resistance), with which transformed bacteria can be selected successfully. Since each bacterium expresses only one plasmid, each successfully transformed bacterium is the carrier of exactly one logomer. Logomers cloned in this manner can then be amplified in a simple manner and in large quantities for further use (on solid or in liquid media or fermenters).

The logomers obtained by Step V of the method of the present invention are cloned for the purpose of isolation and amplification in any cloning vector (plasmid) that carries restriction cut sites that are compatible to the sequence overhangs of the terminators of the logomers (e.g. HindIII, BamHI recognition sequences in the cloning vector pBluescript II KS +/-, Stratagene, Catalogue No. #212207). The cloning is carried out in accordance with the usual laboratory protocols, such as described in [J. Sambrook, E.F. Fritsch, T. Maniatis, *Molecular Cloning, A Laboratory Manual*, (1989)]. For example, the following protocol can be used, and modified in accordance with the knowledge of a person skilled in the art:

a) restriction and preparation of the cloning vector

For cloning, it is possible to use any cloning vectors that are conventionally used for biomolecular processes. (A suitable cloning vector is for example the plasmid pBluescript II KS +/-, Stratagene, Catalogue No. #212207). The plasmid is subjected to a restriction digestion with two restriction enzymes that generate overhangs that are compatible to the terminators of the logomers. For the restriction enzymes, it is possible to use BamHI and HindIII (e.g. by NEB, New England Biolabs, Catalogue No. #136S and #104S). For this, one of the usual restriction protocols can be used, which can be modified in accordance with the knowledge of the person skilled in the art, such as the following:

In a 50 µl preparation, 30 µl plasmid (0.7 µg/µl), 5 µl of 10x buffer (e.g. NEB2, New England Biolabs), 5 µl of BamHl (100 units), 5 µl HindIII (100 units) and 5 µl BSA 10x are incubated for 1-2 hours at 37°C. For control, 1 µl of the preparation as well as the corresponding amount of uncut plasmid is separated electrophoretically on a 1% agarose gel (e.g. UltraPure™ by Gibco BRL, Life Technologies, Catalogue No.: 15510-027).

For the preparation of the DNA, the restriction preparation is phenolized, the DNA is precipitated and taken up again:

- refill preparation to 200 µl with distilled water
- add 200 µl phenole, vortex, and centrifuge at 10000g for 3 minutes
- collect supernatant and add 200 µl chloroform, vortex, and centrifuge at 100000g for 3 minutes
- collect supernatant, acidify with 1/10 Vol. 3M NaAc, and mix with 2.5x volume 100% EtOH, vortex, for at least 15 minutes to -70°C
- centrifuge for at least 15 minutes to 10000g, collect, and discard supernatant
- wash with ca. 500µl of 70% EtOH, centrifuge for 5-10 minutes to 10000g, collect supernatant, and discard
- dry the pellet and take it up again in distilled water, so that the cut plasmid is present at 100 ng/µl to 1 µg/µl (higher concentrations are also possible). The plasmid can be diluted for further ligations if necessary.

b) Ligation of the logomers into the cloning vector

The logomers obtained in Step V of the method of the present invention and the prepared cloning vectors are ligated. The manner of the preparation ensures that, if possible, only the desired ligations take place: The terminators of the logomers carry two different overhang sequences, which are compatible to the overhang sequences of the cloning vector that were created by restriction. Thus, the logomers can ligate only in a predefined direction into the cloning vector. Furthermore, the terminators of the logomers are not phosphorylized, so that the logomers can ligate exclusively with the overhang sequences of the cloning vector. The ligation is carried out in accordance with one of the usual ligations protocols, as described for example in [J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning, A Laboratory Manual, (1989)], which can be modified in accordance with knowledge of a person skilled in the art, for example as follows:

The molar ratio of logomers/cloning vector can be e.g. 500. In 10 µl total volume:

1 µl plasmid (10ng/µl = 5nM)

0.5 µl of 4 µM logomers of Step V of the method of the present invention

1 µl of 10x ligation buffer

0.5 µl of T4 DNA ligase (400u/µl)

7 µl H_2O

are ligated for 12 hours to 16°C.

The protocol be modified in accordance with the knowledge of the person skilled in the art. For example, it is possible to use a protocol for TCL (Temperature Cycle Ligation, [Lund, A.H., Duch, M., Pedersen, F.S, Increased cloning efficiency by temperature-cycle ligation, *Nucleic Acids Research*, **24:(4)**], 800-801, (1996)).

The resulting ligation preparation is now used for the transformation of competent cells, as shown in the example below:

c) transformation of competent cells

The transformation of bacteria (competent cells, host: e.g. dH5α, GIBCO BRL, Catalogue No.: 18258-012) with the ligation preparation of b) is carried out with one of the usual protocols, as described for example in [J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning, A Laboratory Manual, (1989)], which can be modified in accordance with the knowledge of the person skilled in the art, for example as follows:

- 200 µl competent cells (~5x10^7 CFU = colony forming units, stored at -70°C) are thawed and put on ice
- about 1ng of the ligation preparation of b) is pipetted to it and left for 20-30 minutes on the ice, while carefully mixing every 5 minutes
- heat shock: heat the preparation for two minutes to 42°C, then back onto the ice
- add 0.8 ml LB medium (+0.02M $MgSO_4$ +0.01M KCl)
- roll preparation for 20-60 minutes in test tube
- plate 0.1 to 1ml onto agar plate (with antibiotic, e.g. ampicillin) and to 37°C over night.

The transformation functions as a method for selecting successfully cloned logomers and as a method for isolating individual logomers, because each bacterial cell expresses exactly one plasmid.

In addition to the above-mentioned method of cloning, which is preferable in accordance with the present invention, the information-carrying polymers (logomers) obtained in accordance with Step V of the method of the present invention can be isolated and amplified with the following method:

Isolation by hybridization and chromatographic methods

Herein, the molecular mixture resulting from a symbol polymerization is applied on a carrier material, to which oligomers ("anchor sequences") are bonded, which can bond to the terminators of the logomers. The bonded logomers are then taken up individually and can be amplified as needed.

Herein, several methods are applicable:

a) affinity chromatography; herein, anchor sequences that have ends that are compatible to the overhanging ends of the terminators are bonded e.g. to a hydroxylapatite column. The logomers obtained by Step V of the method of the present invention are applied over the column, whereby the logomers can bond to the anchor sequences by hybridization. The logomers bonded in this manner are then taken up individually.

b) anchor sequences are bonded covalently to a membrane (e.g. Gene Screen Plus, DuPont, Biotechnology Systems; Hybond-N, Amersham Life Sciences). The membrane has a grid, which means that it is partitioned into individual fields. Exactly one anchor sequence per field is bonded covalently to the membrane. Then, the logomers obtained in Step V of the method of the present invention are applied onto the membrane and hybridized with the anchor sequences. The membrane is then cut into the individual fields of the grid. The individual fields, which now contain maximally one logomer bonded to the corresponding anchor sequence can now be introduced into separate vessels (such as Eppendorf tubes). Then, the logomers can be separated by denaturing from the membrane. When denaturing, it is important to select a denaturing temperature that it is high enough to separate logomer and anchor sequence, but not so high that the logomer itself melts completely.

Isolation by Dilution

In isolation by dilution, the logomers that are obtained as a mixture of a symbol polymerization are diluted so much, that the target volume statistically contains exactly one molecule. This molecule can then be detected and amplified by PCR. The dilution of the original volume into target volumes can be carried out concurrently, so that an original volume is diluted into n target volumes.

In accordance with the present invention, this is carried out as follows:

For a mixture of logomers, obtained by Step V of the method of the present invention, in an original volume V_a, a dilution factor is determined, that dilutes n logomers contained in V_a such that each target volume V_z statistically contains exactly one logomer. In order to determine the dilution factor, an aliquot of the original volume (e.g. 1 µl) is titrated in a dilution series. Then, for each dilution in the series, an aliquot (e.g. 1 µl) is used as a template in a PCR reaction, in order to determine, in what dilution the logomers are still detectable. Thus obtaining the last dilution containing logomers and the dilution that already contains no logomers, it is possible to specify a rough dilution factor that contains just about one logomer. If necessary, this dilution factor can be determined more precisely by titrating the last dilution that still contains logomers. For this, smaller dilution steps are used (for example, if the first dilution series was diluted at 1 : 10, then the second dilution series can be diluted at 1 : 2; in this manner, the method for precise determination of the dilution factor can be carried out in principle at any desired precision).

Since the PCR of the dilutions is for determining whether there are logomers left in the dilution, the PCR can be carried out in two ways:

a) Two anti-sense primers priming in the terminators, e.g. the 5' strand of the start terminator and the 3' strand of the end terminator, are taken as primers. Otherwise, the PCR conditions are as described below in *Amplification of logomers by PCR*.

b) The PCR is carried out like the PCR for reading out logomers, but a preparation with a pair of primers as described under *Readout of logomers by PCR* below is sufficient.

For control, the DNA fragments obtained by PCR are visualized by gel electrophoresis. For this, e.g. a 2 – 4% agarose gel (e.g. UltraPure™ by Gibco BRL, Life Technologies, Catalogue No.: 15510-027) and as the molecular weight standard e.g. a 50bp ladder (Gibco BRL, Life Technologies, Catalogue No. 10416-014) are used. The resulting gel is colored for 5 minutes in 0.001% ethidium bromide and made visible under UV light.

Amplification of logomers

The logomers obtained by a symbol polymerization can – depending on the intended application – be amplified. This may be necessary in order to visualize the stored information or to further use the logomers as markers for the labeling of substances and objects.

Amplification of logomers by PCR

With this method, logomers are amplified by PCR. The logomer to be amplified serves as a template, and the necessary primers prime in anti-sense in the terminators (either 5' start and 3' end, or 5' end and 3' start), so that the logomer is amplified completely. Alternatively, the primers can also prime outside of the logomers, if the logomer to be amplified is surrounded by further DNA.

For PCR preparations that can be varied in manners known to a person skilled in the art, the following reagents and conditions are used, for example:

dNTPs (e.g. Pharmacia Biotech, Catalogue No.: 27-2035-01/2/3), Taq polymerase (e.g. Gibco-BRL, Catalogue Nos.: 10838-034, 10838-042, 18038-067), PCR buffer, $MgCl_2$ (e.g. GIBCO-BRL, Catalogue No.: 18067-017)

	amount (µl)
dNTP, 10mM	1
PCR buffer 10x	5
$MgCl_2$, 25mM	5
TAQ, 5u/µl	0.5
primer 1, 10µM	1
primer 2, 10µM	1
template (logomer)	1
H_2O	35.5
total volume	50

As the PCR program, it is possible to use the following protocol (primers: 30-mers), for example:

step	action	temperature	duration	goto step	number of repetitions
1	denature	95°C	00:05:00		
2	denature	95°C	00:00:30		
3	annealing	68°C	00:00:30		
4	polymerization	72°C	00:00:30		
5	goto			2	29
6	cool	4°C	(any)		
7	end				

A further object of the present invention are information-carrying polymers that are obtainable by the methods of the present invention described above.

Reading of the information contained in the logomers

Finished logomers can be read either by PCR (polymerase chain reaction), which is preferable in the present invention, or by restriction digestion. The PCR method has the advantage that even smallest amounts of DNA can be amplified such that the logomers can be read directly after the PCR and a gel separation. In the case of binary logomers, the band pattern obtained by PCR on the gel can be read immediately as binary code.

Thus, a further object of the present invention is a method for reading information of information-carrying polymers, that have been obtained and/or isolated and amplified as described above, such that

a) one pair of anti-sense primers each is mixed into n solutions containing the information-carrying polymer, wherein n is the number of oligomers contained as elongators in the polymer;

b) at least n-1 PCR processes are carried out, wherein n is the number of oligomers contained as elongators in the polymer, and one primer of each pair primes in the terminator opposite to the elongator and the other primer primes in the elongator itself;

c) the polymer fragments obtained by PCR are separated according to length using electrophoresis; and

d) the pattern obtained by electrophoresis is read out optically.

The method can be carried out using primers or nucleotides that are fluorescence marked by various colors. Thus, it becomes possible to mark several preparations by different colors and to read several preparations in electrophoretically one track. Furthermore, it is thus possible to automate the readout process with modern sequencing machines (e.g. available by ABI).

Readout of logomers by PCR

In order to visualize the information contained in the logomers, it is possible to read out the logomers by PCR. The method of reading binary patterns by PCR has been described as DNA typing in [Jeffreys, Minisatellite repeat coding as a digital approach to DNA typing, *Nature*, **354**, 204-209, (1991)]. The method described in this article is used to read out logomers in a modified form as *readout PCR*. The differences to the method described in [Jeffreys, Minisatellite repeat coding as a digital approach to DNA typing, *Nature*, **354**, 204-209, (1991)] are that here, synthetically produced templates are used, that do not necessarily contain binary patterns. Furthermore, the PCR and gel electrophoresis conditions are selected such that the result can be read directly from the gel, without having to further amplify the signals obtained as band patterns by blotting.

a) PCR

To read a logomer, at least n – 1, but usually n PCR preparations are needed for n elongators of the grammar. Every PCR preparation contains the logomers to be read as a template and furthermore a pair of anti-sense primers, of which one primes in the elongator, and the other one in the terminator opposite to the elongator (e.g. 5'-"start" and 3'-0, see Fig. 5). For the PCR, it is possible to use any commercial thermal cycler (e.g. PTC-100, MJ Research). A length of 20-30bp has been found to be suitable as the length of the primers, but it is also possible to use other lengths. The annealing temperature has to be selected in accordance with the length of the primers (and can be determined e.g. using the program Oligo 5.0). Suitable annealing temperatures for 20-mers are about 55°C, and for 30-mers about 65°C to 74°C.

For the PCR preparations, the following reagents and conditions are used, for example that can be varied in manners known to the person skilled in the art: dNTPs (e.g. Pharmacia Biotech, Catalogue No.: 27-2035-01/2/3), Taq polymerase (e.g. Gibco-BRL, Catalogue Nos.: 10838-034, 10838-042, 18038-067), PCR buffer, $MgCl_2$ (e.g. GIBCO-BRL, Catalogue No.: 18067-017)

	amount (µl)
dNTP, 10mM	1
PCR buffer 10x	5
$MgCl_2$, 25mM	5
TAQ, 5u/µl	0.5
primer 1, 10µM	1
primer 2, 10µM	1
template (logomer)	1
H_2O	35.5
total volume	50

In order to obtain enough DNA to read the resulting band pattern after the gel electrophoresis directly from the gel, it is possible to prepare each PCR preparation m times. For example, m = 4.

As the PCR program, it is possible to use the following protocol (primers: 30-mers), for example:

step	action	temperature	duration	goto step	number of repetitions
1	denature	95°C	00:05:00		
2	denature	95°C	00:00:30		
3	annealing	69.5°C	00:00:30		
4	polymerization	72°C	00:00:30		
5	goto			2	29
6	cool	4°C	(any)		
7	end				

After the PCR has been carried out, it may be advantageous for the readability of the band pattern of the subsequent gel electrophoresis to hold as much amplified DNA in as little volume loaded onto the gel as possible. The volumes can be narrowed down by any suitable method (e.g. with a SpeedVac, e.g. SpeedVac Concentrator, Savant)

b) gel electrophoresis

The DNA fragments obtained from the PCR for each elongator have different lengths. Separating them by electrophoresis, the various length fragments result in a specific pattern, by which it is possible to identify the logomers to be read (see Fig. 5 and Fig. 6). For the gel electrophoresis, it is possible to use a 4% agarose gel (e.g. UltraPure™ by Gibco BRL, Life Technologies, Catalogue No.: 15510-027), and as the molecular weight standard it is possible to use e.g. a 50bp ladder (Gibco BRL, Life Technologies, Catalogue No.: 10416-014). The separation of the DNA is performed in a gel chamber in any suitable manner, such as 1:45 hours at 60V. However, the parameters can be selected freely, in particular, it is possible to shorten the gel run-time.

The resulting gel is colored for five minutes in 0.001% ethidium bromide and is visualized under UV light.

An example of the resulting gel is shown in Fig. 6.

Reading out logomers by restriction digestion

Logomers can be read out by restriction digestion. For this, the elongators must be configured such that they carry asymmetrically shifted restriction cut sites.

Each elongator carries a specific restriction cut site. For example, the restriction enzyme EcoRV (e.g. NEB, Catalogue No. #195S) can be used for 0-elongators, and SmaI (e.g. NEB, Catalogue No. #141S) can be used for 1-elongators.

For the readout, the logomer to be read out is cut in different restriction preparations. For this, one restriction preparation per elongator is formed with the corresponding restriction enzyme. The DNA fragments obtained from the restriction have different lengths. When they are separated by electrophoresis, the different length fragments result in a specific pattern, by which it is possible to identify the logomer to be read out.

The following example shows the DNA fragment lengths of all binary logomers with 4 bit without terminators at an elongator length = 30bp; restriction cut site per elongator after 10bp.

	0	1	
binary number			
0000	10, 40, 40, 40, 30	160	
0001	10, 40, 40, 70	130, 30	
0010	10, 40, 80, 30	90, 70	*
0011	10, 40, 110	90, 40, 30	
0100	10, 80, 40, 30	50, 110	*
0101	10, 80, 70	50, 80, 30	
0110	10, 120, 30	50, 40, 70	
0111	10, 150	50, 40, 40, 30	
1000	50, 40, 40, 30	10, 150	
1001	50, 40, 70	10, 120, 30	
1010	50, 80, 30	10, 80, 70	
1011	50, 110	10, 80, 40, 30	*
1100	90, 40, 30	10, 40, 110	
1101	90, 70	10, 40, 80, 30	*
1110	130, 30	10, 40, 40, 70	
1111	160	10, 40, 40, 40, 30	

In the above example (of 4 bits), it can be seen that the length pattern of the numbers 0010 and 0100 for restriction of the 0-bit is ambiguous. However, it is possible to distinguish between 0010 and 0100 using the length pattern for restriction of the 1-bit.

In particular, the above example was carried out as follows, although the protocols can be varied according to the knowledge of a person skilled in the art:

a) DNA extraction

A bacterial colony obtained e.g. by *Isolation and amplification of logomers by cloning* is used to inoculate 2-5ml of a 37°C overnight culture (e.g. LB medium with 10 µg/l ampicillin, as described in [J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning, A Laboratory Manual, (1989)]. As the cloning vector, it is possible to use for example pGEM-T Easy (Promega, Catalogue No. A1360) and binary logomers without terminators. The plasmid DNA containing the logomers is isolated from the overnight culture (for example using the Qiagen Plasmid Miniprep Kit, Qiagen, Catalogue No. 12123, or by alkaline lysis as described in [J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning, A Laboratory Manual, (1989)].

b) Restriction

The plasmid DNA obtained from a) is cut for n elongators into n restriction preparations. Each restriction preparation contains a restriction enzyme that cuts in a specific elongator and a restriction enzyme that the logomer cuts from the cloning site. In this example, the following preparations were used:

	name	quantity	volume (µl)
plasmid DNA	logomer in pGEM-T easy (see above)	500 ng/µl	6
buffer	NEB2	10x	1
BSA		10x	1
enzyme 1	ECO RV	10 u/µl	1
enzyme 2	ECO RI	10 u/µl	1
total			10

	name	quantity	volume (µl)
plasmid DNA	logomer in pGEM-T easy (see above)	500 ng/µl	6
buffer	NEB4	10x	1
BSA		10x	0
enzyme 1	Sma I	10 u/µl	1
enzyme 2	ECO RI	10 u/µl	1
H_2O			1
total			10

The listed preparations are incubated for one hour at 37°C.

c) Gel electrophoresis

The DNA fragments obtained by b) are separated electrophoretically (e.g. with 4% Agarose UltraPure™ by Gibco BRL, Life Technologies, Catalogue No.: 15510-027), colored with ethidium bromide (0.001%) and made visible under UV light (see Fig. 8).

The above-described methods enable for example the implementation of binary logomers:

Binary logomers are generated by grammars with:

$\Sigma := \{0, 1, s_0, s_1, s_2, \ldots, s_{n-1}, s_n, e_0, e_1, e_2, \ldots, e_{m-1}, e_m\}$ with n, m $\in \mathbb{N}$, and

n, m $\geqq$ 0.

They allow the simplest and most universal representation of data, which is also applied to conventional data processing machines. Binary logomers are the result of a bit polymerization. Since practically any symbols and rules can be coded depending on the chosen grammar, it is also possible to generate any data and data types with them.

The above-described methods enable for example the implementation of character alphabets:

Logomers can be used to code the characters of a chosen alphabet. Different character lengths can be chosen, but it is advantageous to use character lengths that are also used in conventional data processing machines (half byte, 1 byte, 2 bytes, 4 bytes, 8 bytes, etc.). Also the conventions for interpreting the represented characters can be selected freely. The characters can be numbers, letters, alphanumeric characters or any other data structure.

The above-described methods enable for example the implementation of a 1-byte alphabet:

Given is a grammar G = ($\sum$, V, R, S) with the terminal alphabet $\sum := \{0, 1, s, e\}$, set of variables V := $\{S_0, S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8\}$, start symbol S and a set of rules

R :=

{

$S := sS_0$

$S_0 \rightarrow 0S_1$

$S_0 \rightarrow 1S_1$

$S_1 \rightarrow 0S_2$

$S_1 \rightarrow 1S_2$

$S_2 \rightarrow 0S_3$

$S_2 \rightarrow 1S_3$

$S_3 \rightarrow 0S_4$

$S_3 \rightarrow 1S_4$

$S_4 \rightarrow 0S_5$

$S_4 \rightarrow 1S_5$

$S_5 \rightarrow 0S_6$

$S_5 \rightarrow 1S_6$

$S_6 \rightarrow 0S_7$

$S_6 \rightarrow 1S_7$

$S_7 \rightarrow 0S_8$

$S_7 \rightarrow 1S_8$

$S8 \rightarrow e$

}

wherein:

s := start

e := end.

Example:

With the rules according to the given grammar, all 8-bit characters can be generated, for example: generation of 8-bit 0, 00000000:

$S \rightarrow sS_0 \rightarrow s0S_1 \rightarrow s00S_2 \rightarrow s000S_3 \rightarrow s0000S_4 \rightarrow s00000S_5 \rightarrow s000000S_6 \rightarrow s0000000S_7 \rightarrow s00000000S_8 \rightarrow s00000000e$

for example: generation of 8-bit 255, 11111111:

$S \rightarrow sS_0 \rightarrow s1S_1 \rightarrow s11S_2 \rightarrow s111S_3 \rightarrow s1111S_4 \rightarrow s11111S_5 \rightarrow s111111S_6 \rightarrow s1111111S_7 \rightarrow s11111111S_8 \rightarrow s11111111e$

for example: generation of 8-bit 85, 01010101:

$$S \rightarrow sS_0 \rightarrow s0S_1 \rightarrow s01S_2 \rightarrow s010S_3 \rightarrow s0101S_4 \rightarrow s01010S_5 \rightarrow s010101S_6 \rightarrow s0101010S_7 \rightarrow s01010101S_8 \rightarrow s01010101e$$

The sequences necessary for the in vitro implementation are produced according to the NFR method, as described in Steps II and III. Algomers and logomers are manufactured as described in the inventive Steps IV and V. The contained logomers are preferably isolated and amplified by the method described under *Isolation and amplification of logomers by cloning*. The alphanumeric symbols generated in this manner can now be used for various technical purposes (e.g. marking and identifying substances), and, possibly after additional technical steps, read out in accordance with the inventive method for reading information from information-carrying polymers (as described above).

Since the generated binary patterns can be read, as desired, as data or symbols, they can be interpreted as alphanumeric symbols, for example. If they are interpreted as numbers, then the stated grammar generates 8-bit random numbers.

From a sufficiently large set of random 8-bit binary patterns, it is possible to isolate all 8-bit patterns and set up as a library (e.g. for representing all alphanumeric characters).

The above-described methods enable for example the representation of character strings:

Since the algomers can be produced such that logomers of different length can be generated, it is also possible to represent character strings. However, for practical reasons, it is preferable to provide only few characters per logomer. For example, a single logomer can contain 4 bits (half-byte) or 8 bits (1 byte). (If more bits are necessary, then it is preferable to use only multiples of 8 bits = 1 byte.) In binary representation, it is then possible to represent any alphanumeric character in any coding (e.g. ASCII, ANSI, ISO 8859-x, Unicode.) One character can also span several logomers (for example: half-byte representation, in which a 1-byte character spans two logomers, or Unicode, in which a 16-bit character spans two 8-bit or four 4-bit logomers.

For the representation of character strings, it is then necessary to provide the individual characters with positional information. For that, it is sufficient to use for that grammar different terminators (e.g. different "start" terminators), whose sequences represent the corresponding positional information.

Example:

Given is a grammar $G = (\sum, V, R, S)$ with a terminal alphabet $\sum := \{0, 1, e, s_0, s_1, s_2, \ldots, s_{n-1}, s_n\}$, a set of variables $V := \{S_0, S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8\}$, a start symbol S and a set of rules

R :=

{

$S := s_i S_0$

$S_0 \rightarrow 0S_1$

$S_0 \rightarrow 1S_1$

$S_1 \rightarrow 0S_2$

$S_1 \rightarrow 1S_2$

$S_2 \rightarrow 0S_3$

$S_2 \rightarrow 1S_3$

$S_3 \rightarrow 0S_4$

$S_3 \rightarrow 1S_4$

$S_4 \rightarrow 0S_5$

$S_4 \rightarrow 1S_5$

$S_5 \rightarrow 0S_6$

$S_5 \rightarrow 1S_6$

$S_6 \rightarrow 0S_7$

$S_6 \rightarrow 1S_7$

$S_7 \rightarrow 0S_8$

$S_7 \rightarrow 1S_8$

$S_8 \rightarrow e$

}

wherein:

s_i := start with i = {0, ..., n}

e := end.

The 1-byte alphabet generated with this grammar is sufficient, for example, for all alphanumeric characters of the ASCII, ANSI or ISO 8859-x standards. For a character string of length n, n different terminators are necessary, so that the character strings generated with this grammar are constructed as follows:

$s_0$xe, $s_1$xe, ..., s_{n-1}xe, $s_n$xe

wherein x is any binary representation with, in this case, 8 bits.

In this case:

$s_0$xe := character x at position 0 (0^{th} character)

$s_1$xe := character x at position 1 (1^{st} character)

$s_2$xe := character x at position 2 (2^{nd} character)

etc.

With this grammar, it is possible to represent for example the character string “Elisabeth” in ASCII code:

$s_0$01000101e, $s_1$01101100e, $s_2$01101001e, $s_3$01110011e, $s_4$01100001e, $s_5$01100010e, $s_6$01100101e, $s_7$01110100e, $s_8$01101000e.

Alternatively, it is also possible to select $\sum$:= {0, 1, s, e_0, e_1, e_2, ..., e_{n-1}, e_n}, whereby the character string “Elisabeth” in ASCII code is represented as:

s01000101e_0, s01101100e_1, s01101001e_2, s01110011e_3, s01100001e_4, s01100010e_5, s01100101e_6, s01110100e_7, s01101000e_8.

A further possibility is $\sum$:= {0, 1, s_0, s_1, s_2, ..., s_{n-1}, s_n, e_0, e_1, e_2, ..., e_{n-1}, e_n}, in which case the character string “Elisabeth” in ASCII code is represented as:

$s_0$01000101e_0, $s_1$01101100e_1, $s_2$01101001e_2, $s_3$01110011e_3, $s_4$01100001e_4, $s_5$01100010e_5, $s_6$01100101e_6, $s_7$01110100e_7, $s_8$01101000e_8.

In half-byte representation with positional information, the character string “Elisabeth” in ASCII code is represented as:

$s_0$0100e, $s_1$0101e, $s_2$0110e, $s_3$1100e, $s_4$0110e, $s_5$1001e, $s_6$0111e, $s_7$0011e, $s_8$0110e, $s_9$0001e, s_{10}0110e, s_{11}0010e, s_{12}0110e, s_{13}0101e, s_{14}0111e, s_{15}0100e, s_{16}0110e, s_{17}1000e.

Because of the positional information, which is represented by the sequence of terminators, the individual characters can be processed independently from one another, and read independently from one another, for example.

With such an alphabet, the representation of any data type is possible. For example, it is possible to combine two bytes each (0^{th} + 1^{st}, 2^{nd} + 3^{rd}, ..., n^{th} + $n+1^{th}$) to a 2-byte code (e.g. Unicode). Alphanumeric character strings can be used to represent any indicator , name, numbers, date, etc.

The sequences necessary for the implementation in vitro are produced according to the NFR method, as described in Steps II and III. Algomers and logomers are produced as described in the inventive Steps IV – V. The resulting logomers are preferably isolated and amplified by the method described under *Isolation and amplification of logomers by cloning*. The alphanumeric symbols generated in this manner can now be used for various technical purposes (e.g. marking and identifying substances), and, possibly after additional technical steps, read out in accordance with the inventive method for reading information from information-carrying polymers (as described above).

The above-described methods enable for example the implementation of a random number generator:

For certain problems in computer science (e.g. simulations) and mathematics, random numbers are required. Random numbers are ordinarily generated by computer-based algorithms. However, since those algorithms are deterministic, the generated random numbers are only pseudo-random numbers. In addition, the generated number series are randomized at varying quality, due to the varying quality of the algorithms.

For some applications, however, real random numbers are required. In that case, a physical process has to be included in the generation of random numbers. This process can be for example the noise of the sound card in a PC, which is then processed by corresponding software.

Using the methods described in this specification, it is possible to implement a real random number generator, which generates random numbers much faster than is possible with conventional methods.

The random number generator is implemented with the following grammar (grammar for binary random numbers of any length):

G = (∑, V, R, S) with ∑ := {0, 1, s, e}, V := {S}

R :=

{

S := sA

A → 0A

A → 1A

A → e

}

wherein

s := start

e := end

With this grammar, it is possible to form words with the above-mentioned terminal alphabet

∑ := {0, 1, s, e}

All words of the language L begin with an "s" and end with an "e", and a random number (including 0) of random bits ("0" and "1") is contained therebetween.

Example: Words of the language L, that can be formed according to R:

S → sA → se

S → sA → s1A → s1e

S → sA → s0A → s00A → s001A → s0010e

S → sA → s1A → s10A → s100A → s1000A → s10000A → s100000A → s1000000e

The binary patterns generated as words of the language can be read as letters, numbers, alphanumeric characters, or character strings. When reading them as the binary representation of numbers, then the above binary patterns are, in decimal representation:

Ø (empty)

1

2

32

and can be used as random numbers. In principle, any of the sequences is suitable for in vitro implementation, as long as they are unambiguous and maximally dissimilar to another.

For example, it is possible to use the following algomers:

sA:

agctttatatctccatttgccctagtgaag
    aatatagaggtaaacgggatcacttcaacc

A ➔ 0A:

ttggcgagatatcaacgccaccccttgctt
    gctctatagttgcggtggggaacgaaaacc

A ➔ 1A:

ttggcgacccgggaaacaactattgctagt
    gctgggccctttgttgataacgatcaaacc

A ➔ e:

ttggtgcgggagttggaagcaactacgatg
    acgccctcaaccttcgttgatgctacctag

The necessary sequences are commercially available. They can, for example, be ordered as 40nmol, PAGE-purified (ARK-Scientific, Darmstadt) and taken up by distilled water (recommended storage at –20° C). Algomers are produced from these sequences in the manner described in Step IV of the present invention. The algomers are polymerized to logomers, as described in Step V of the present invention, and can be used to read out information from information-carrying polymers (as described above). Random numbers generated like this are shown in Fig. 6.

The above-described methods enable for example the encryption of logomers:

The information contained in logomers can be encrypted. For this, at least one of the primers needed for the readout PCR functions as the secret key (see Fig. 11). It is preferable that this is one of the primers priming in the terminators. If the sequence of this primer (key sequence) and the primer itself are known only to authorized personnel, then the information contained in the encrypted logomers is also accessible only by the authorized personnel.

In order to prevent the information contained in the logomers from being read out without that key, further precautions are necessary. Possible attempts to break the encryption could be based on reading the logomers by amplifying non-secret sequences. If bacterial vectors are used for the cloning of logomers, a potential attacker could try, for example, to amplify the entire cloning site by PCR and then read it by sequencing, or to "read in" the resistance genes necessary for the selection into the cloning site. Also, the elongators themselves might serve as attack points for PCR-based reading of the secret sequence(s).

What these attacks have in common is that they might try to decrypt the key sequence with non-secret sequences, and thus read the encrypted sequence. A defense against such attacks is possible by using only completely secret sequences for the information transport via logomers. However, this might be too labor-intensive and expensive. Instead, or supplementarily, it is possible to mix the information-carrying logomers with sequences (dummy sequences) that contain the same potential attack points (e.g. bits, resistance genes), but each containing different key sequences. With these dummy sequences, potential attacks are successfully blocked. For unauthorized accesses, all individual sequences are indistinguishable and therefore the encrypted information remains concealed. The more dummy sequences are mixed into a logomer, the more difficult it becomes to break the key sequence.

This method corresponds to molecular stenography and is illustrated by the following example of the encryption of a character of a 1-byte alphabet:

The used grammar is $G = (\sum, V, R, S)$ with terminal alphabet $\sum := \{0, 1, s_{key}, s_0, s_1, s_2, \ldots, s_{f-1}, s_f, e\}$, wherein $f \in \mathbb{N}$, $f \geqq 1$. s_{key} is the secret key. For a key length of l, the maximal theoretical encryption is $Key_{max} = a^{l}$, the maximally usable encryption $Key_{eff} = a^{l-d}$ wherein d is the number of bases, in which the dummy key has to be different from the correct key, so that in the readout PCR no dummy key reads the correct sequence and conversely, the correct key cannot read out dummy logomer (but only the target logomer). For highly specified PCR conditions, it is possible to use $d \geqq 1$. Furthermore, Key_{imp} = number of dummy keys + 1 is the actually used encryption, and $Key_{min} = 1 + x$ is the minimum encryption. Here, x is the number of dummy logomers necessary for a minimum encryption. In the case of using n logomers as information carriers, x can be selected to be larger than n.

For character strings, there is automatically an additional encryption, because the *order* of the characters, which is not known to the potential attacker, acts as an additional encryption.

Example: The following grammar G is used for a grammar with character strings of n 1-byte characters:

The used grammar is $G = (\sum, V, R, S)$ with terminal alphabet $\sum := \{0, 1, s_{key}, s_{key0}, s_{key1}, s_{key2}, \ldots, s_{keyn-1}, s_{keyn}, s_0, s_1, s_2, \ldots, s_{f-1}, s_f, e\}$, wherein $n, f \in I\!N$, $n \geqq 1$. n expresses the number of used real keys. f is the number of used dummy keys; the used set of variables V and the used set of rules R can be selected freely and depending on the information to be encoded.

Grammars with different character coding function analogously.

The sequences necessary for the implementation in vitro are produced according to the NFR method, as described in Steps II and III. Algomers and logomers are produced as described in the inventive Steps IV and V, and the contained logomers are preferably isolated and amplified by the method described under *Isolation and amplification of logomers by cloning*. The logomers generated in this manner can be read out with the readout PCR described above, wherein – as described above – one of the primers necessary for readout, preferably one of the primers priming in the terminators, functions as the secret key. With the grammar described above, “dummy sequences” (“dummy logomers”) are generated in addition to the information-carrying logomers, which are supposed to prevent unauthorized readout access to the information contained in the logomers. The method can be used for all applications of logomers and has, among others, the advantage to be very effective due to the specificity of the primer used as the secret key.

Based on the above-described method, it is possible to implement symmetric and asymmetric encryption schemes. For a symmetric encryption scheme, it is sufficient if the communicating participants A and B agree on a secret key sequence. Participant B encrypts a message stored as a monomer by manufacturing logomers only with the secret terminators and mixing additional logomers that carry other terminators into the message. Only A is then in position to provide the primer pair that is necessary for readout, because only A knows the necessary terminator sequence.

An asymmetric encryption scheme, on the other hand, necessitates additional effort, such as the application of molecules with irregular forms (see for example Fig. 12 and Fig. 13): When a communication partner B wants to send an encrypted message to A, then B receives a public key from A, which B uses to encrypt the message represented by a logomer. The public key A includes a pair of terminators and a pool of additional dummy terminators with similar properties, but deviating sequences and other 3'-overhang sequences. Of the genuine terminators, only the 3' overhang sequence is known, with which it can be linked to the – publicly known – elongators. To encrypt a message, B generates logomers, using the key that was publicly provided by A (and which includes terminators and dummy terminators) for the symbol polymerization reaction. This configuration of the terminators ensures that only A is in a position to read out the thusly encrypted message: Since only A knows the actual sequence of the genuine terminators, only A can read, by readout PCR, the message that has been encrypted as logomers.

Since the public key can theoretically be attacked via the known overhang sequence to the elongators (because the dummy terminators may not possess this overhang sequence, a potential attacker can link any sequence to the genuine terminators, whereby the sequence of the terminators can be identified), molecules based on irregular shapes, such as based on the Y-shaped molecules shown in Fig. 12, are used for the terminators. The illustrated Y-shaped molecules can be combined further to tree-like branching molecules. The root of such a terminator molecule realized as a tree then contains the overhang sequence compatible to the elongators, whereas only one of the branches of the tree contains the genuine terminator sequence. Since, as explained above, additional dummy logomers are mixed into the encrypted message, only A, who knows the genuine terminator sequence, can filter the proper message from the set of molecules by linking with the compatible vector, whereas a potential attacker who does not know the genuine terminator sequence cannot do this.

The above-described methods enable for example encryption with real-time identification:

Without reading out the information contained in encrypted logomers, the presence of the key sequence itself can be indicative of whether a marked product is genuine or not. The information about the authenticity of the marked product can be verified or falsified in a process in which the key primer serves as hybridization probe of a fluorescence detection reaction.

A further object of the present invention is therefore the use of information-carrying polymers, in particular of the information-carrying polymers obtained in accordance with the present invention, for the encryption of information.

The above-described methods enable for example a test for the quality of oligonucleotides

A typical problem regarding the production of oligonucleotides, for example for molecular biology, is the quality control of the oligonucleotides, which tends to be problematic because of the production process and the varying quality of the used chemicals.

The method described the Steps I - V of the present invention can be used to test the quality of oligonucleotides. For this, a binary grammar such as for the random number generator (see above) or a unary grammar such as for the production of molecular weight standards (see below) is used for the generation of logomers of any length. Under otherwise identical conditions, the length of the logomers obtained by a symbol polymerization (Step V of the method of the present invention) is then directly proportional to the quality of the used oligonucleotides: The longer the electrophoretically visualized logomers are, the better is the quality of the used oligonucleotides. An example of the ladder of logomers of different length obtained with a symbol polymerization is shown in Fig. 3.

A further object of the present invention is therefore the use of information-carrying polymers, in particular of the information-carrying polymers obtained in accordance with the present invention, for the quality control of synthetically manufactured oligonucleotides.

The above-described methods enable for example the manufacture of logomers as markers and signatures

Because of their ability to represent, in principle, any piece of information, the logomers described here can be used as markers for labeling manufactured goods, products, substances, and devices.

A further object of the present invention is therefore the use of information-carrying polymers, in particular of the information-carrying polymers obtained in accordance with the present invention, as markers or signatures.

The logomers can be used as a "binary label," that is added to a product, and that contains information about the thusly labeled product. The logomers can contain e.g. information about

a) manufacturer
b) product (e.g. serial number)
c) application purpose
d) substance class
e) danger class
f) quality
g) purity
h) production and expiration date

It is possible to provide substantially any product or manufactured goods with substantially any information. Since the logomers are non-toxic and biodegradable, they can also be used for highly sensitive products, such as food, medical and pharmaceutical products.

Labeled products can be for example:

a) chemical products, e.g. varnishes, paints, oils, lubricants, fuels, solvents, inks, etc.

b) liquid materials: solutions, suspensions, emulsions .

c) genetically engineered products

d) food, e.g. for the labeling of genetically altered ingredients or for monitoring of foodstuffs during the production process

e) medical and pharmaceutical products

f) paper products, documents, money

g) devices

There is a need to label products in a lot of different fields and due to different requirements. Reasons for labeling can be for example: quality assurance during the production process, supervising product purity and avoiding contamination, preventing counterfeiting and product piracy, detection of certain product components, and product labeling with the desired production information. The need to label also exists in sensitive areas, such as the labeling of food, medical and pharmaceutical products and genetically engineered or modified products. Here, it is desirable that information about the product is available directly on the product, in particular when it is desirable to avoid mix-ups or counterfeits and also to identify different components of a product.

There are numerous examples for the necessity to label products. For example, logomers can be used as serial numbers that are mixed into car paint, so that in the case of an insured event (e.g. accident or theft) it is possible to identify the car owner. By product labeling, it is possible to supervise the quality and purity of chemical products and avoid contaminations. A permanent problem is the counterfeiting of documents, signatures and money, which necessitates highly developed labeling methods.

In particular with animal diseases and epidemics (such as BSE, swine fever, salmonellosis, etc.) in mind, there is the problem of quality assurance of foodstuffs and the labeling of end products, in order to keep contaminated products away from consumption. A further problem is food that contains genetically engineered ingredients. These ingredients cannot, or only with tremendous difficulty, be detected in the end product. This problem is aggravated by the fact that many foodstuffs contain ingredients of various origins and the production routes are often unclear.

Labeling is also a problem in the field of blood products, in which contamination may occur by pooling. Also in this field, it is desirable to have information about the identity, origin, production date and expiration date available directly on the product. This is also true for medical and pharmaceutical products.

Conventionally, various carbon compounds, polyanilines, liquid crystals or other chemical compounds are used to label and mark colors, paints, fuels, etc. The used substances and compounds have various deficiencies: For example, some of them are toxic, hardly biodegradable, the available information capacity is severely limited, they interact chemically with certain substances or are expensive and their production is labor-intensive.

For the use of labeling sensitive and highly sensitive products such as food and drugs, the above-mentioned substances and compounds are not suitable at all, due to the mentioned deficiencies.

The requirements of a method for labeling any product include at least:

- The label should be available directly in or at the product;
- the label should be easily detectable;
- the label must be harmless to health.

In contrast to conventional compounds, the logomers described herein have the following advantages:

- coding of any information,
- high storage capacity,
- harmless to health,
- chemically, biologically, pharmacologically and genetically neutral,
- easily biodegradable (biomolecules),
- no environmental harm,
- easy detection,
- high compatibility to existing techniques of computer science and molecular biology (PCR),

- information can be encrypted and are also suitable for authentication and encryption,
- identification of individual ingredients in product mixtures,
- cheap to produce in large quantities (cloning, amplification with bacteria in fermenters).

These properties make logomers much better suited for the mentioned purposes than the conventional techniques. Moreover, application fields are opened, that have been closed to conventional technologies.

Because of their properties, nucleic acid based logomers can also be used for labeling particularly sensitive products such as food and pharmaceutical products.

The reasons for this are grounded in the chemical nature of nucleic acid based logomers on the one hand, and in the coding of information in logomers on the other hand:

1) As the "basic building blocks of life," nucleic acids are elementary building blocks of all known biological life forms. Thus, they cannot be harmful to any known biological organism due to their *chemical* nature ("biochemical compatibility"). However, due to the contained genetic *information* (of the contained program), nucleic acids may very well be potentially harmful when they are read by an organism: If they encode for example toxic proteins or if (as in the case of viral genes) they can "reprogram" the target organism. However, due to the following reasons, the logomers used here cannot contain information that is harmful to an organism.

2) Logomers do not contain genetic, i.e. biologically relevant information ("semantic incompatibility"): Regarding the contained information, logomers correspond to letters, numbers or series of letters and numbers in binary coding, that are readable to us or a computer, but not for the genetic apparatus of an organism. For a biological organism, they are simply "nonsense code".

3) Logomers, and in particular the algomers from which they are assembled, are too short even to contain *accidentally* biologically relevant information.

4) In order to preclude any possibility (i.e. the unlikely event that the binary

information of the logomers is in some way interpreted by an organism as genetically relevant information), the sequences used for the coding are provided with stop codons in the case of sensitive application fields, so that they can never be read by a biological organism.

5) As elementary components of all biological organisms, nucleic acids are ubiquitous in the environment. They are degraded within an extremely short time.

In order to be used as a marker, logomers can be generated according to the Steps I – V of the present invention, and amplified in accordance with the methods of the present invention as described above. Depending on the application, it is possible to use different methods and different regenerative processes to amplify the logomers. In order to label petrochemical products for example, it is possible to amplify the logomers by cloning in bacteria, for example. To avoid unnecessary contamination, the resulting bacteria can be lysed. A special cleaning of the logomer DNA is usually not necessary.

To manufacture logomers for sensitive and highly sensitive products, however, a more elaborate amplification and cleaning method is necessary. Its goal is to obtain maximally pure logomers. To avoid biological intermediate steps, the amplification can be carried out by PCR, as described in accordance with the present invention. If amplification by cloning is carried out, then the cleaning of the resulting DNA of bacteria is necessary. If the logomers are cloned in bacterial vectors, then a targeted degradation of these resistance genes (e.g. by restriction enzymes) is part of any further processing.

Logomers can be connected immediately to the product to be labeled. The logomers can be mixed into liquids, for example, or applied to solids. In the case of genetically engineered products, the logomers can be introduced into the product to be labeled by recombination or cloning techniques.

The labeling of substances or products by logomers enables among others the monitoring of production processes, quality control, the identification of individual components and the quantitative and qualitative detection of contaminations.

An aspect of the present invention is explained in the following, with an example of manufacturing computer-compatible data (bytes/multibytes) of nucleic acids and their use in marking various substances. However, the present invention is not limited to this example.

1. A coding of bytes and multibytes in form of nucleic acids is defined.
2. Suitable sequences for the representation of bytes and multibytes are constructed.
3. A library of byte-representing nucleic acids is manufactured in vitro.
4. Bytes are linked to multibytes.
5. Biochips for the identification of bytes and multibytes are manufactured.
6. Different (organic and inorganic) materials and objects as well as individual genes are labeled (and, if necessary, encrypted) by bytes or multibytes.
7. The information of labeled materials, objects or individual genes is read out (and, if necessary, decrypted).

Explanation of 1: The representation of data with nucleic acids is supposed to ensure that the data are computer-compatible. For this purpose, the representation by bytes and multibytes is selected, as it is the most universal representation of computer data.

For this, a grammar G_{C32} is defined, with $G_{C32} = (\sum, V, R, S)$, terminal alphabet $\sum := \{x_m, o_m, s_m, e_m\}$, variable alphabet $V := \{A, B, C\}$, start symbol S and set of rules $R := \{S := oA, A \rightarrow s_mB, B \rightarrow Ce_m, C \rightarrow x_n\}$ wherein n, $m \in \mathbb{N}$, $n = \{0, 1, \ldots, 255\}$ and $m \geqq 0$. The grammar describes the construction of molecules of the form osxe, wherein x can take on any byte value between 0 and 255, s describes the byte position, and o and e are provided for the concatenation of individual bytes into multibytes (see Fig. 29 and Fig. 30). To simplify the manufacturing process, restriction sites are used as variables.

In addition, the used nucleic acids must have logical, physical, chemical, and biological properties. In particular they should be easy to amplify, easy to read, and easy to interact with biological sequences in a predefined manner.

For this purpose, the data-representing nucleic acids are defined such that

a) they represent bytes

b) they can also represent more complex data structures (multibytes)

c) individual bytes can be linked into multibytes

d) they can be read with a biochip

e) they can be used for the manufacture of 1-byte and multibyte biochips

f) individual bytes and multibytes can be amplified by PCR

g) they are clonable

h) they can be linked with other nucleic acids so as to mark them

i) they carry as few restriction sites as possible

a) To let the data-representing nucleic acids represent bytes, exactly 256 unambiguous sequences must represent all values of a byte. For this, the molecular bytes contain an unambiguous sequence x, of which there are 256 different alleles. These alleles are referred to as x_0 to x_{255} and enumerate exactly all byte values ($x_0 = 1$, $x_1 = 1$, ..., $x_{255} = 255$) (see Fig. 25, Fig. 26, Fig. 27, Fig. 28 and sequence listing).

b) In order to be able to represent more complex data structures (multibytes), such as 32-bit numbers and strings, the byte molecules are provided with position information. This information is contained in the sequence s (see Fig. 25, Fig. 26, Fig. 27, Fig. 28 and sequence listing). Each position is represented by an unambiguous sequence s_i. All byte positions are then enumerated by the different alleles of s (s0 = position 0, s1 = position 1, ... etc.).

c) Although for the representation of multibytes, it is in principle sufficient to distribute the data to be represented over independent bytes with the corresponding position information (as multi-strand multibytes: for example, to provide a paint with a serial number of four bytes length, it is possible to mix four individual bytes, each having exactly one position information s_0 to s_3, independently into the paint), it is sometimes preferable to link bytes to single-stranded multibytes. For this purpose, the bytes carry the sequences o and e (see Figs. 25 to 28 and sequence listing), with which they can be linked by ligation or overlap assembly and PCR to single-stranded multibytes (see Fig.

29 and Fig. 30). These single-stranded multibytes can then be used as markers. In the case of parallel overlap assembly, the individual bytes are cut out with an appropriate restriction enzyme (here: EcoRV, see Fig. 28).

d) Data-representing nucleic acids can be read with sequencers, for example. However, in order to read them as easily and fast as possible, it is suitable to read them using biochips. For this purpose, biochips are manufactured that include all 256 different x-sequences in single-strand form (and thus all byte values) in ascending order (see Fig. 32 and Fig. 33). Biochips with this configuration are also an object of the present invention. The reading of a certain byte value x_i is carried out by denaturing the corresponding data molecule to single strands and hybridization with the chip. The hybridized sequences are marked, in correspondence to the typical operation principles of biochips, for example by fluorescence marking. After the hybridization of data molecules with a 1-byte chip, exactly one marked position on the biochip is obtained, which thus marks exactly one byte value (see Fig. 36). For example, a data molecule that carries the sequence x_{192} hybridizes exactly at position 192 on the chip, and thus it can be identified by its position on the chip.

e) Since the principle of reading the data molecules with biochips is based on the hybridization of complementary sequences, the bytes are configured such that they can also be used for the manufacturing of the corresponding biochips. For this, the byte molecules carry enzymatic restriction sites, so that from one complete byte molecule either the sub-sequence containing x or the sub-sequence containing s and x can be cut out. These sub-sequences are then denatured and spotted onto a chip carrier. In order to make a 1-byte chip, all 256 x-sequences are applied in ascending order on the chip. Thus, it is possible to identify a sequence with previously unknown x by its hybridization position. When only the x sub-sequence is used to make the 1-byte chip, then so-called X-chips are obtained (see Fig. 32), and if the sub-sequence including s and x is used, then so-called SX-chips are obtained (see Fig. 33). X- and

SX-chips can be combined to form multibyte chips. A further condition for the manufacture of biochips is the availability of sufficient amounts of sequences. To ensure this, the byte molecules can be made directly with oligosynthesizers, they can be cloned, or they can be amplified from few template molecules by PCR.

f) In order to explicitly amplify individual bytes (e.g. from single-stranded multibytes), the sub-sequences referred to as o, s and e can serve as priming sites (see Fig. 25, Fig. 26, Fig. 27 and Fig. 28).

g) In order to be clonable, the byte molecules carry sticky ends that are compatible to restriction sites (e.g. Xhol and Xbal, see Fig. 28).

h) In order to link bytes and multibytes more easily with the desired nucleic acids, it is possible to append further terminators (adaptors) to bytes and multibytes by ligation or by overlap assembly and PCR (see Fig. 29 and Fig. 30). These adaptors carry restriction sites or recombination sites, so that they can be linked to other nucleic acids either by restriction and ligation or by recombination. This is in particular advantageous for the marking of nucleic acids, in particular genes.

i) In the context of linking bytes and multibytes with other nucleic acids, it is preferable that in particular multibytes carry only few restriction sites, so that the user of the nucleic acid that is provided with the bytes and multibytes can work with them as usual and does not have to do without any restriction sites, which otherwise destroy the bytes. Therefore, the bytes are configured such that after the manufacture of single-stranded multibytes, of all restriction sites that are based on sequence palindromes of length 6 only the restriction sites Afl1l, Nhel, and Ncol are contained in the multibytes and therefore cannot be used like that, for example, for nucleic acids marked with multibytes. If necessary, it is even possible to remove the remaining restriction sites from the bytes and multibytes without losing the byte and position information, by amplifying multibytes by PCR and mutant primers such that the restriction sites are blocked.

Explanation of 2: The sequences necessary for implementation are synthesized using the NFR method described above. The sequences are configured such that all sequences are as dissimilar to one another as possible. The x sub-sequences have a GC content of 50%, and the o, s, and e sequences have a GC content of 66%, multibytes contain three sequence palindromes of lengths 6, which are recognized one each by Afll1, Nhel and Ncol.

Explanation of 3: The byte molecules can be manufactured directly using an oligosynthesizer. In order to manufacture larger amounts of byte molecules, it is however advantageous to devise a clone library containing all byte molecules.

Explanation of 4: Due to the included byte position information, bytes can represent multibytes without being directly linked to one another (logical linking to multi-stranded multibytes). However, it is advantageous for some applications, such as the labeling of nucleic acid strands, to link the individual bytes physically together to one multibyte (single-stranded multibyte, see Fig. 30 and Fig. 31). To accomplish this, the individual bytes are linked to one another either by ligation or by overlap assembly and PCR.

Explanation of 5: Biochips for reading bytes and multibytes are manufactured with sub-units of the molecules that function as bytes. In principle, there are at least two architectures: the X-chip (see Fig. 32), which is manufactured with x-sub-sequences (see Fig. 25 to Fig. 28), and the SX-chip (see Fig. 33), which is manufactured with sx-sub-sequences (see Fig. 25 to Fig. 28). Both chips are 1-byte chips and can be combined to multibyte chips. Their operation principle is similar: The reading of information from data molecules is carried out by hybridization with the chip. The byte value of a data molecule can be determined by its hybridization position.

Thus, it is also possible to use multibyte chips as data storages (ROM and RAM): The hybridization of the chip with a specific data molecule corresponds to assigning a value to a storage cell. The storage cell can then be erased by denaturation (e.g. by temperature increase or change of pH). In addition, it is also possible to provide specific sequences that are hybridized with the chip with optically active molecules of different colors, so that specific color patterns can be generated and the multibyte chips can be used as a display. Combining sequences of different melting temperature with optically active molecules of different colors, a color display in accordance with a thermometer can be devised.

The difference between the X-chip and the SX-chip is that the SX-chip can also read (and store) byte positions in addition to the byte value. The advantage of the X-chip is that very large multibyte arrays can be made of identical X-chips. The advantage of the SX-chip is that it can read single-stranded multibyte molecules, without having to hybridize the individual bytes separately.

Explanation of 6: With the adaptors L and R (see Fig. 29 and Fig. 30), it is possible to connect sequences with bytes and multibytes serving the purpose to connect them with nucleic acids. The adaptors contain for example suitable restriction sites or recombination sites. For example, a multibyte can be cloned in a plasmid using suitable restriction sites, whereby the plasmid receives a label. This label can be for example a 32-bit serial number (see Fig. 31). Another application example is providing a data molecule with recombination sites, wherein the data molecule is introduced for example into non-coding regions (e.g. introns) of genes, thus labeling these.

Explanation of 7: The information in substances, objects and molecules that have been labeled by data molecules can be read out by extracting it from the labeled substance, cleaning it if necessary, amplifying it if necessary by PCR and then sequencing it or identifying it by hybridization with a byte or multibyte chip, as described above.

The methods described above thus enable the unambiguous labeling of organic and inorganic substances and objects as well as of nucleic acid constructs and genes.

The methods described above also enable the manufacturing of data storages and optical displays: To store computer data, multibyte arrays are used. The writing of a single byte is carried out by hybridizing one 1-byte chip with a nucleic acid (that is optically labeled, if necessary), that contains a defined x-sequence exactly once, that x-sequence corresponding to the value to be written (e.g. x_{192}). The reading of the data is performed like the reading out of a biological DNA array, e.g. by scanning. The erasing of data is carried out by denaturation and, if necessary, subsequent cleaning.

The above-described methods enable for example the labeling of single molecules, nucleic acids and genetically engineered or modified products:

The logomers described herein can be used to label genetically engineered or modified products.

Conventionally, "classical" biomolecular methods such as PCR, restriction digestion, Southern blot hybridization, and sequencing have been used to identify genetically engineered and modified products in food. However, with these methods, a characterization of genetically engineered products and ingredients is very labor-intensive and also necessitates separate methods and processes for each product to be characterized.

All in all, there is so far no universal labeling method for genetically engineered products. Such a labeling method must fulfill several criteria:

- the label must be absolutely harmless to health,
- the label should be inseparably linked to the labeled product,
- the label should be forgery-proof,
- the label should be able to store enough information,
- the label should be detectable and readable in small amounts,

The use of nucleic acid based logomers to label genetically engineered products satisfies these criteria and furthermore has the following advantages:

- logomers can store practically any information;
- individual ingredients can be detected fast and with high sensitivity;
- a method that is standardized in detail can be applied to read out the information (readout method of the present invention, for which standardized primers can be used);
- the sequence of the searched genes can be completely unknown;
- the sequence of the detected genes can be secret, without compromising identification or authentication;
- additional security mechanisms can be implemented additionally by encryption.

In particular, the labeling of genetically engineered or modified products is accomplished for example as follows:

To manufacture markers, any suitable grammar is used, e.g. the one for the random number generator, the 1-byte alphabet, or for character strings as described above. The characters needed to represent the markers are taken from the characters generated with the grammar and added to the product to be labeled. The characters manufactured as logomers can be introduced in the following ways in the product to be labeled:

a) by admixture; herein, the logomers are manufactured, isolated, and amplified, as described in the method of the present invention.

b) by cloning, as described in the method of the present invention; Fig. 6 shows for example the labeling of bacterial clones by logomers, which have been produced by using the grammar for the random number generator described above.

c) by restriction and ligation; herein the logomers are manufactured, isolated and amplified, in accordance with the present invention. The logomers obtained in this manner are then connected by restriction and ligation with the nucleic acid to be labeled. For this, the terminators (see Fig. 2) and adaptors (see Fig. 30) described above can carry the desired restriction sites.

d) By using recombinative techniques; herein the logomers are manufactured, isolated and amplified, in accordance with the present invention. The logomers obtained in this manner are then introduced by recombinative techniques into the product to be labeled. For this, it is possible to use for example the methods of gene targeting by homologous recombination [Capecchi M.R., Altering the genome by homologous recombination. Science, 244(4910), 1288-1292, (1989)] or e.g. the Cre-IoxP System [Kilby, N.J., Snaith, M.R. & Murray, J.A., Site-specific recombinases: tools for genome engineering, *Trends Genet.*, **9**, 413-421, (1993)]. For example, logomers obtained in accordance with the present invention can be set before or behind a certain nucleic acid sequence (gene), that they are supposed to label. In this case, the logomer is “tacked” as a “tag” to a genetically engineered product and can provide information about manufacturer, product group, danger class, expiration date etc. (see Fig. 15).

The method is in particular suitable for the labeling of single molecules, in particular single nucleic acids, and preferably single genes.

The method is applicable to all organisms (microorganisms, plants, animals) For example, clones of microorganisms or plants can be labeled. The method is also suitable to label food and to label cattle in order to fight the cattle epidemic BSE.

The above-described methods enable for example the labeling of food:

Food can be unhealthful or even dangerous because of diseases, epidemics, or contamination. Examples are diseases such as BSE, swine fever and salmonellosis. It is desirable to make it possible to identify product and manufacturer, so that in an emergency, certain products or product series can be immediately taken from the market, examined, and proper counter-measures can be taken.

For highly sensitive products such as food, the criteria for labeling genetically engineered products have to be applied stringently (see: *The above-described methods enable for example the labeling of single molecules, nucleic acids and genetically engineered or modified products*).

All in all, there is also no universal labeling method for genetically engineered food products. Conventionally, the identification of genetically engineered or modified ingredients in food products is therefore difficult and labor-intensive. For such labeling, “classical” biomolecular methods such as PCR, restriction digestion, Southern blot hybridization, and sequencing are used to characterize the products.

In the wake of the cattle epidemic BSE, an immunological labeling method for the labeling especially of cattle and cattle produce (i.e. beef and milk) has been suggested, which is based on an immunization of the animals to be marked with specific proteins (“Ear mark in blood traces down BSE cattle”, Süddeutsche Zeitung Nr. 283, 08.12.1998, page v2/9). With the production of specific antibodies caused by this immunization in the marked target animal, which is supposedly detectable in any kind of tissue, the label can be read with an immuno-assay (ELISA). However, such a method is on principle limited to higher vertebrates. Moreover, the available amount of information for this method is highly limited, the information-carrying proteins are not heat-resistant, and the method requires comparatively large amounts of substances for immunization and detection reaction.

As explained in the section *The above-described methods enable for example the labeling of single molecules, nucleic acids and genetically engineered or modified products*, logomers based on nucleic acids can also be used to for labeling food products. The logomers can contain information about the manufacturer, expiration date, and serial number, among others. The method for labeling food can be the same as described in *The above-described methods enable for example the labeling of single molecules, nucleic acids and genetically engineered or modified products*. Using the genetic recombination and cloning techniques listed in that section, single organisms can be labeled such that a single organism and all its descendants are identifiable.

The method for labeling by logomers thus enables:

- monitoring of the target product throughout the entire production and processing procedure down to the end user,
- a unified, fast and simple identification of individual ingredients.

Logomers based on nucleic acids are also suitable as an admixture to label beverages.

The above-described methods enable for example the labeling of medical and pharmaceutical products

Because of their non-toxicity, logomers based on nucleic acids are suitable for labeling sensitive and highly sensitive products. Besides foodstuffs, medical and pharmaceutical products can be labeled. In order to preclude mix-ups, contamination, and counterfeits, logomers can be admixed for example to medical and pharmaceutical products.

The advantage of labeling with logomers is that the label is connected immediately to the labeled product, and products can be labeled individually, for example. In this manner, a monitoring of the production process is possible, products can be identified even after refilling them several times, and contamination as well as the pooling of probes can be detected. For example in the case of blood storage, it is possible to label production and expiration date, blood type, and manufacturer using logomers.

The above-described methods enable for example authentication and copy protection

Logomers can be used for authenticating products, objects and devices. For this, the logomers are preferably generated and amplified in accordance with the present invention.

If logomers obtained in this manner are applied or admixed to products, objects, and devices, then the information contained in the logomers can be read out in accordance with the present invention at any time. If the generation of the logomers was performed using encryption techniques, as described above, then the read-out necessitates authorized access.

For example, logomers obtained in accordance with the present invention and dissolved in an aqueous solution (preferably of 10^3 to 10^{18} molecules/µl) can be applied immediately on documents for labeling. As such, they can serve as a certificate of authenticity of a document marked that way. To read out the logomers serving as the label, a small shred of the paper (1mm^2 is sufficient) is used as a template of a PCR reading reaction in accordance with the present invention. An example of this is shown in Fig. 16. In the read-out experiment of this figure, a logomer in an aqueous solution of ca. 10^9 molecules/µ. was dried for about one hour on 3M PostIt paper and 1 mm^2 of the PostIt was used as a template of a readout PCR for proof of principle. It is possible to use documents of any paper quality, such as 80 g/m^2 standard copying paper.

The same method is also suitable for labeling money or bills, which is advantageous, because:

- bills that have already been printed can be marked,
- it is possible to assign serial numbers,
- encryption can be used.

In order to test the authenticity of documents, if possible in real-time, it is possible to use one of the primers used for readout in the PCR as a hybridization probe for a subsequent coloring detection reaction (e.g. with an intercalating colorant, such .as ethidium bromide).

Another example is the authentication of liquid solutions, suspensions and emulsions with logomers. Here, it is possible to mark inks individually with logomers, so that the authenticity of signatures can be tested.

Another example is the labeling of automobiles by mixing logomers into the car paint. For this, logomers are added as an admixture to the car paint (or another component of the car). The added logomers can contain information about the serial number, manufacturer, car type, or the like. Based on this information, a vehicle can be identified in the case of an accident, theft, or illegal disposal. An advantage of this method is that already traces of the car paint are sufficient to identify the vehicle, which is particularly desirable for accidents. Another advantage is that it becomes very labor-intensive to fake the identity of a vehicle. To do this, first, all labeled components would have to be removed, and second, a fake identity would have to be established. This is already very difficult because of the complicated mechanical procedure, and moreover, the use of cryptographic methods can make the copying or mimicking of serial numbers as difficult as desired. To read out the information contained as logomers, the logomers are isolated from the car paint, and can then be read out by the method for the reading of information contained in logomers described above, preferably by PCR. For this, a sample of the paint is obtained, dissolved in a suitable solvent (such as turpentine) and the same volume of water is added. The following centrifugation separates hydrophilic and hydrophobic parts. Then, taking the aqueous phase (containing the hydrophilic nucleic acid based logomers), the logomers contained in the aqueous phase are precipitated by the regular methods (e.g. as described in [J. Sambrook, E.F. Fritsch, T. Maniatis, *Molecular Cloning, A Laboratory Manual*, (1989)]. After the precipitation, the resulting logomers can be read out with the methods according to the present invention, such as by readout PCR.

Thus, other objects of the present invention are the use of information-carrying polymers, in particular information-carrying polymers obtained in accordance with the present invention, for the purpose of quality assurance, copy protection, labeling of food, labeling of genetically engineered, chemical, medical, and pharmaceutical products, the labeling of organisms, the labeling of documents, the labeling of money, the labeling of objects and machinery, the labeling of solutions, suspensions and emulsions, as well as the authentication of persons and objects.

The above-described methods enable for example the production of a molecular weight standard

One of the typical methods of biomolecular analysis is the electrophoretic longitudinal separation of DNA and RNA fragments. Usually, in order to determine the length of such fragments in electrophoresis, a molecular weight standard ("ladder") is used, that contains fragments of known length and that is used for the comparison with fragments of the samples to be analyzed.

So far, there are molecular weight standards that contain either irregular ladders (e.g. Boehringer V, Boehringer Mannheim, Catalogue No.: 821705) or regular ladders with regular fragment-length distances (e.g. 50bp ladder Gibco BRL, Life Technologies, Catalogue No.: 10416-014). None of the conventional ladders has length distances that are shorter than 10bp (e.g. 10bp ladder Gibco BRL, Life Technologies, Catalogue No.: 10821-015).

With the methods described herein, it is possible to manufacture specifically DNA fragments of different length and predefined length intervals. These fragments can be used as molecular weight standards, and the method makes it possible to realize, in principle, any length and length distances. In particular, length distances up down to 1bp interval, i.e. molecular weight standards of the highest resolution, are possible.

The method is based on using logomers for the template of a readout PCR, whose result is a mixture of DNA fragments of predefined length and predefined length distances. It is based on the inventive method for generating and reading logomers.

DNA fragments of various lengths in very short, regular intervals can be obtained when a unary polymer functions as a template for a readout PCR with nested primers. The 5' primer primes at the start sequence of the polymer. The anti-sense 3' primer is attained by a group of primers, that starts nested in the elongator (shown in Fig. 17 as three levels of primers shifted with respect to one another). Since the polymer is made of repetitions of only one elongator molecule, n*m bands are obtained for n repetitions and m nested 3' primers. This method also works laterally reversed with a primer priming in the end terminator and corresponding anti-sense 5' primers. The resulting bands can be used as molecular weight standards, e.g. in electrophoresis.

For high resolution molecular weight standards, the use of unary logomers and nested primers is most preferable. However, the method is not limited to this, because in principle, all grammars described in Step I of the present invention can be used.

Unary logomers of any length can be produced with a grammar G = (∑, V, R, S) with a terminal alphabet ∑ := {0, s, e}, a set of variables V := {A}, a start symbol S and a set of rules R

R :=
{
  S := sA
  a ➔ aA
  A ➔ e
}

If it is necessary to produce unary logomers of a predefined length, then the set of variables and thus the number of elongators must be chosen to be larger (such as for example in the grammar for the implementation of a 1-byte alphabet described above). Depending on the desired length of the DNA fragments to be generated, it is also possible to vary the length of the algomers.

The generated logomers can be isolated and amplified with the methods of the present invention.

The generation of DNA fragments of predefined length and predefined length intervals is carried out preferably by PCR, similar to the inventive method for reading out logomers. For high resolution molecular weight standards, at least one primer, but mostly a number of primers is used, which are shifted against one another. The number of primers per elongator depends on the desired length intervals. If it is desired to use algomers of a length of 30bp and if the lengths of the generated DNA fragments should have a difference of 10bp, then three primers, that are shifted exactly 10bp with respect to one another, are used.

For the synthesis of sequences for producing algomers and logomers, the requirements described for Step II of the method in accordance with the present invention have to be fulfilled. In particular, because the partial sequences serving as the template have the same GC content, the NFR method ensures the use of nested primers and the balance of the AT/GC ratio of the generated DNA fragments. This last aspect is crucial, regarding the fact that the running behavior of the DNA fragments in the electrophoresis depends not only on length, but also on the configuration of the fragments. This is particularly true for high resolution molecular weight standards, in which there are only small length intervals.

A further object of the present invention is thus the use of information-carrying polymers, in particular information-carrying polymers obtained in accordance with the present invention, for the manufacture of molecular weight standards.

The above-described methods enable for example the manufacturing of a polymeric data storage

The logomers explained above can be used as RAM (read and write) or ROM (read only) data storages with high capacity and low energy consumption In order to ensure the highest possible information density and to make the handling of the storage as easy as possible, the logomers are bonded to a solid carrier. This makes sure that the logomers can be addressed individually, i.e. written, read, and erased individually (see Fig. 9).

The necessary write and read operations are carried out on the basis of the enzymatic reactions described for Step V of the present invention and in the section *Reading of the information contained in the logomers*, which are controlled by changing the temperature. The temperature changes can be carried out by laser or by heating and cooling the solid carrier, e.g. in a thermal cycler.

To write, a modification of the above-described method of symbol polymerization is used. The polymerization is carried out on a solid carrier, and the algomers are concatenated with one another by repeated restriction-ligation cycles.

In order to manufacture single, addressable storage cells, "anchor molecules" are irreversibly connected (by UV irradiation or in a furnace) to the carrier in certain intervals. Thus, logomers that can be removed in an erasure process can be written reversibly onto the carrier provided with such anchor molecules.

For the write process, first, specific algomers ("starter algomers") are bonded to the anchor molecules by hybridization. These starter algomers are the point of origin of a symbol polymerization, in which further algomers are concatenated with one another. Different from Step V of the method of the present invention, the concatenation of algomers is carried out by repeated cycles of restriction and ligations. In each cycle, the last algomer of a polymerizing chain is cut with a restriction enzyme, so that a single-stranded overhang sequence (elongation point) is created, which can then be bonded by ligation to the subsequent algomer.

For this write process, the algomers must be provided with a specific design:

a) Each algomer to be written has a single-stranded overhang sequence, with which it can bond to the elongation point of a logomer.

b) An algomer that has only a single-stranded overhang sequence with which it can bond to the elongation point, but not a restriction cut site, by which it can be recognized by restriction enzymes and thus serve as an elongation point, is referred to as "end algomer."

c) Every algomer to be written that is not an end algomer has a double-stranded end containing the recognition sequence for the selected restriction enzyme. At a restriction, the enzyme splits the algomer, so that a single-stranded overhang sequence is created, which then can serve as an elongation point.

d) Every finished logomer has exactly one start algomer and one end algomer. Start and end algomers are, in accordance with the definition given above, *terminators*.

e) The overhang sequences of the algomers are compatible to one corresponding selected restriction enzyme, so that an algomer to be concatenated can bond to the elongation point of a logomer and the elongation point created subsequently by restriction is identical to its predecessor.

f) The sequence following the overhang sequence in the algomer is selected such that a concatenated algomer cannot be split off again by a subsequent restriction process.

Suitable solid carriers include membranes, such as Gene Screen Plus (DuPont, Biotechnology Systems, Catalogue No.: NEF-986 or NEF-987), Hybond-N (Amersham

Life Sciences, Catalogue No.: RPN 82N or RPN 137N), silicon or silicate surfaces or glass. The anchor molecules are covalently bonded to the solid carrier (by UV irradiation or heat).

In order to make the molecules individually addressable, they have to be arranged on the carrier in regular intervals.

Suitable enzymes include commercially available enzymes, preferably enzymes that are not deactivated at 65°C (e.g. HindIII).

The writing operations are based on the fact that the enzymes used for writing have different temperature optimums. For example, the temperature optimum of the ligase used for the concatenation of symbols is 16°C, whereas the necessitated restriction enzyme functions only at 37°C. Thus, the timed writing of symbols in cycles of restriction and ligation is possible.

Since all read and write operations are carried out using enzymes, it is necessary to be able to adjust the temperature of the carrier. For this, all storage cells can be subjected simultaneously to the same operation by using a thermal cycler and material that can be manipulated thermally. Alternatively, storage cells can be accessed individually and independently from one another, if a correspondingly small write/read head is used. This write/read head must serve as a pipetting device for the necessitated molecules (enzymes, algomers) on the one hand, but must also generate the temperature needed for a certain manipulation (e.g. with a laser).

In the erasure operation, a logomer is detached from its anchor molecule by denaturation. The anchor molecule is then available again for a write operation. For a read operation, a logomer is detached from an anchor molecule by denaturation. Then, it can be read with the methods of the present invention.

The write, erase, and read operations described above are carried out by enzymes at various temperatures. For this, using a material that can be manipulated thermally, all storage cells can be subjected to the same operation. Alternatively, storage cells can be accessed individually and independently from one another, if a correspondingly small write/read head is used. This write/read head must serve as a pipetting device on the one hand, but must also generate the temperature needed for a certain manipulation.

The advantage of the polymer storage explained above is in its higher storage density (DNA molecules have a size of around 10^{-10}m) as compared to conventional materials, and its much lower energy consumption (regarding the energy needed for enzymatic reactions, see [L.M. Adleman, Molecular Computation of Solutions to Combinatorial Problems, *Science*, **266**, 1021-1024, (1994)]). Moreover, the environmental compatibility is improved considerably, because the polymers described herein are completely non-toxic.

Compared to the aqueous solutions used conventionally in DNA computing, the approach described above has the advantage of higher storage densities and addressability of individual polymers.

Thus, another object of the present invention is to provide a polymeric data storage containing information-carrying polymers according to the present invention, methods for linking molecules in accordance with the present invention, and methods for reading or methods for isolating and amplifying in accordance with the present invention.

The above-described methods enable for example the manufacturing of a DNA computer

Using the above-described information representation by algomers and logomers, it is possible to configure a DNA computer. The DNA computer includes (see Fig. 10):

a) oligonucleotide synthesizers

b) thermal cyclers

c) pipetting device

d) device for isolating polymers

e) device for separating nucleic acids, such as an electrophoresis system

f) scanner

g) control computer (work station, e.g. PC)

The DNA computer is configured as a hybrid system, in which simple, calculation-intensive algorithms or the storage of data are executed with DNA and the necessary devices (devices a) to e)), and a conventional computer (preferably a work station, e.g. a PC) which is used as host and control computer.

The DNA computer contains at least one thermal cycler serving as an "information reactor," which is controlled with the control computer (e.g. MJ Research PTC-100 Eppendorf Mastercycler). In the tubes or microtiter plates of the thermal cycler, there are molecule mixtures of nucleic acids, enzymes and further chemical substances (such as buffers and nucleotides) that are needed for the algomer assembly (Step IV of the method of the present invention) and the symbol polymerization (Step V of the method of the present invention). The DNA computer is programmed by the implementation of algorithms as grammars (Step I of the method of the present invention). The monomer sequences needed for the grammars are produced with the NFR method and using an oligonucleotide synthesizer (e.g. ABI 392, ABI 398, ABI 3948 by Perkin-Elmer Applied Biosystems) (Steps II and III of the method of the present invention). These monomer sequences are introduced as input into the thermal cycler, in which the algomer assembly (Step IV of the method of the present invention) is carried out.

The resulting algomers are introduced in one or more reaction chambers of the thermal cyclers, where they are linked to logomers (symbol polymerization, Step V of the method of the present invention). The resulting logomers are isolated and amplified in a separate device. This device can operate in accordance with the principles that are described for the method of the present invention. After isolation and amplification, the logomers can be read. For this, they are read out in accordance with the present invention in a thermal cycler. The nucleic acid fragments resulting from the readout process can be read by gel electrophoresis, as described above. Here, the gel electrophoresis device serves as the output device of the DNA computer, and the resulting band patterns are read into the control computer by a scanner.

The control computer now undertakes further calculations and controls based on the obtained results, and, if necessary, causes further molecular reactions. For example, based on the obtained results, the control computer can cause the new implementation of grammars (manufacture of oligomers with the NFR processes). In this manner, the information obtained by molecular processes as logomers can serve not only as passive data, but also as instructions (and addresses). Thus, a self-regulating process of different algorithms, which is necessary for a universal data processing, is possible.

The DNA computer processes information as oligomers and polymers. Cloning vectors (plasmids) that are located in a liquid solution or on a solid, addressable carrier serve as storage cells of the data.

The DNA computer as described above is programmable by grammars and can be controlled by a conventional computer, which functions as a control computer and host system.

Since the DNA computer can carry out several operations in molar orders of magnitude (for example the symbol polymerization, as described in Step V of the method of the present invention, in $10^{12} - 10^{20}$ elementary operations per second), a possible application is the calculation of simple, calculation-intensive algorithms. However, the computer can also be used for other applications, such as the generation of certain, regular DNA structures, which are of interest in nano-technology, for example.

A further object of the present invention is to provide a DNA computer including information-carrying polymers in accordance with the present invention. A further object of the present invention is to provide a DNA computer, in which a reading method in accordance with the present invention and/or an isolation and amplification method in accordance with the present invention are used.

The above-described methods enable for example the manufacturing of smallest molecular structures with logomers

It is possible to manufacture the smallest molecular structures (nanostructures) with the logomers described herein. Components of such structures are algomers, which can be assembled to ordered patterns of logomers. For example, various algomers can be doped with different impurity molecules (ligands), in order to obtain ordered, higher molecular patterns of the corresponding ligands. Herein, the logomers function as the "skeleton" of molecular components of ligands.

For example, a binary logomer can contain an alternating pattern of conducting and non-conducting ligands. By arranging many logomers on a solid carrier, it is possible to produce conductors in the nanometer order.

But it is also possible to use, for example, biomolecules, antibodies, optically active molecules as ligands.

Logomers can be used as an "intelligent" glue (see Fig. 18). For this, the hybridization of complementary nucleotides is taken advantage of. When logomers are applied to surfaces to be glued together, and the logomers are bonded for example covalently to the surfaces to be bonded, since only surfaces with complementary nucleic acids form hydrogen bonds then otherwise identical surfaces can adhere to one another if they have complementary sequences. The "intelligence" of the glue is thus grounded in the highly selective adhesion effect, which depends on the sequences used. It is also possible to precisely adjust the adhesion strength for sequences to be glued together: First, it is possible to adjust the adhesion strength of surfaces by adjusting the density of logomers on them, and second, it is possible to vary the melting point of the logomers via the AT/GC ratio, so that surfaces glued together loose their adhesion at different temperatures. Due to the fact that harmless, easily degradable biomolecules are used for the glue, it is also possible to use such glues for medical applications (e.g. in surgery, or microsurgery). Another application is the use of the above-described method in high-precision applications and authentication applications.

If sequences for the binding of ligands are added to the algomers used to produce the logomers, then it is possible to attain a stronger adhesive effect. Such ligands can be, in the case of DNA, DNA-binding proteins, such as specific antibodies directed at the DNA-sequences. These can then be further bonded among one another with a further ligand for example (see Fig. 19). A simple, non-programmable adhesive effect can also be attained without logomers, if the proteins used (e.g. antibodies) can bond to the surface to be glued (such as antibodies, if the surface to be bonded is their antigen) (See Fig. 20). Simpler glues are possible by genetically producing proteins bonding specifically to certain surfaces.

An alternative method is to connect the surface to be glued together not by weak interaction but by covalent binding. Different from the method described above, the adhesion force is then adjusted with the density of algomers. For this, the surface to be glued together are provided with algomers that carry sticky ends that are complementary to one another. These can then hybridize and be covalently bonded with each another by ligation.

With logomers, it is possible to produce surfaces of very different structure and different physical properties. What these applications have in common, is that with the logomers, almost any desired pattern can be formed in the nanometer order, so that they can serve as the skeleton of the smallest molecular components.

A further object of the present invention is thus the use of information-carrying polymers, in particular of information-carrying polymers obtained in accordance with the present invention, to produce or process smallest molecular structures, or as molecular-scale adhesive.

The above-described methods enable for example the manufacturing of nanotechnological construction systems.

With the NFR method and the "parallel extension" method described above, it is possible, for the first time, to translate larger grammars into molecules. These grammars are not limited to regular grammars. Due to the reusability of sequences, it is for example also possible to program Ψ-molecules [Matthias Scheffler, Axel Dorenbeck, Stefan Jordan, Michael Wüstefeld, Günter von Kiedrowski, Self-Assembly of Trisoligonucleotidyls: The Case for Nano-Acetylene and Nano-Cyclobutadiene, *Angewandte Chemie Int. Ed.*, **38**(22), 3312-3315, (1999)] and DX-molecules [Eric, Winfree, Furong Liu, Lisa A. Wenzler & Nadrian C. Seeman, Design and self-assembly of two-dimensional DNA crystals, *Nature*, **394**, 539-544, (1998)].

A further object of the present invention is therefore the use of the NFR-method, the "parallel extension" method and the translation of grammars in molecules to produce components of nanotechnological construction systems.

The above-described methods enable for example the controlled, programmable production of biologically active nucleic acids.

For this, biologically active sequences such as restriction sites, recombination sites, centromers, promoters, exons, genes, etc. are used as terminals instead of artificial sequences to represent symbols. The biologically active sequences can then be assembled in a controlled, programmable manner to biologically active constructs, such as genes or artificial chromosomes.

A further object of the present invention is thus the use of information-carrying polymers, in particular information-carrying polymers in accordance with the present invention, for the controlled production of biologically active nucleic acids.

References

Patents:

International Publication Number	Inventor	Title
US 4683202	Kary B. Mullis, Kensington California	Process for amplifying nucleic acid sequences
WO 97/07440	Adleman, Leonhard, M; 18262 Hastings Way, Northridge, CA 91326 (US)	MOLECULAR COMPUTER
WO 97/29117	GUARNIERI, Frank [US/US]; 62 Lake Street, Brooklyn, NY 11223 (US). BANCROFT, Frank, Carter [US/US]; 51 Dewey Street, Huntington Station, NY 11743 (US).	A DNA-BASED COMPUTER
US 5804373	Schweitzer; Allan Lee, Plainsboro, NJ; Smith; Warren D., Plainsboro, NJ	Molecular automata utilizing single- or double-stranded oligonucleotides

Publications:

L. M. Adleman, Molecular Computation of Solutions to Combinatorial Problems, Science, **266**, 1021-1024, (1994)

Breslauer, K.J., Frank, R., Blocker, H., Marky, L.A., Proc. Natl. Acad. Sci., **83**, 3746-3750, 1989

Capecchi M.R., Altering the genome by homologous recombination, *Science*, **244** (4910), 1288-1292, (1989) Chomsky, N., Three models for the description of language, *JACM*, **2:3**, 113-124, (1959)

Chomsky, N., On certain formal properties of grammars, *Inf. and Control*, **2:2**, 137-167, (1959)

Chomsky, N., Formal properties of grammars, *Handbook of Math. Psych.*, **2**, 323-418, (1963)

Frank Guarnieri, Makiko Fliss, Carter Bancroft, Making DNA Add, *Science*, **273**, 220-223 (1996)

John E. Hopcroft, Formal Languages And Their Relation To Automata, (1969)

Jeffreys, Minisatellite repeat coding as a digital approach to DNA typing, *Nature*, **354**, 204-209, (1991)

Kilby, N. J., Snaith, M. R. & Murray, J. A., Site-specific recombinases: tools for genome engineering, *Trends Genet.*, **9**, 413-421, (1993)

Rolf Knippers, Molekulare Genetik, *Georg Thieme Verlag*, (1997)

Benjamin Lewin, Genes V, *Oxford University Press*, (1994)

Lund, A.H., Duch, M., Pedersen, F.S, Increased cloning efficiency by temperature-cycle ligation, *Nucleic Acids Research*, **24:(4)**, 800-801, (1996)

J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning, A Laboratory Manual, (1989) Jens Niehaus, DNA Computing: Assessment and Simulation, *Master Thesis Paper at the Faculty of Computer Science of Dortmund University, Department XI*, 116-123, (1998)

Qi Ouyang, Peter D. Kaplan, Shumao Liu, Albert Libchaber, DNA Solution of the Maximal Clique Problem, *Science*, **278**, 446-449, (1997)

Eric Winfree, Xiaoping Yang, Nadrian C. Seeman, Universal Computation via Self-assembly of DNA: Some Theory and Experiments, *Proceedings of the 2nd DIMACS Meeting on DNA Based Computers, Princeton University, June 20-12*, (1996)

Eric, Winfree, Furong Liu, Lisa A. Wenzler & Nadrian C. Seeman, Design and self-assembly of two-dimensional DNA crystals, *Nature*, **394**, 539-544, (1998)

Matthias Scheffler, Axel Dorenbeck, Stefan Jordan, Michael Wüstefeld, Günter von Kiedrowski, Self-Assembly of Trisoligonucleotidyls: The Case for Nano-Acetylene and Nano-Cyclobutadiene, *Angewandte Chemie Int. Ed.*, **38**(22), 3312-3315, (1999)